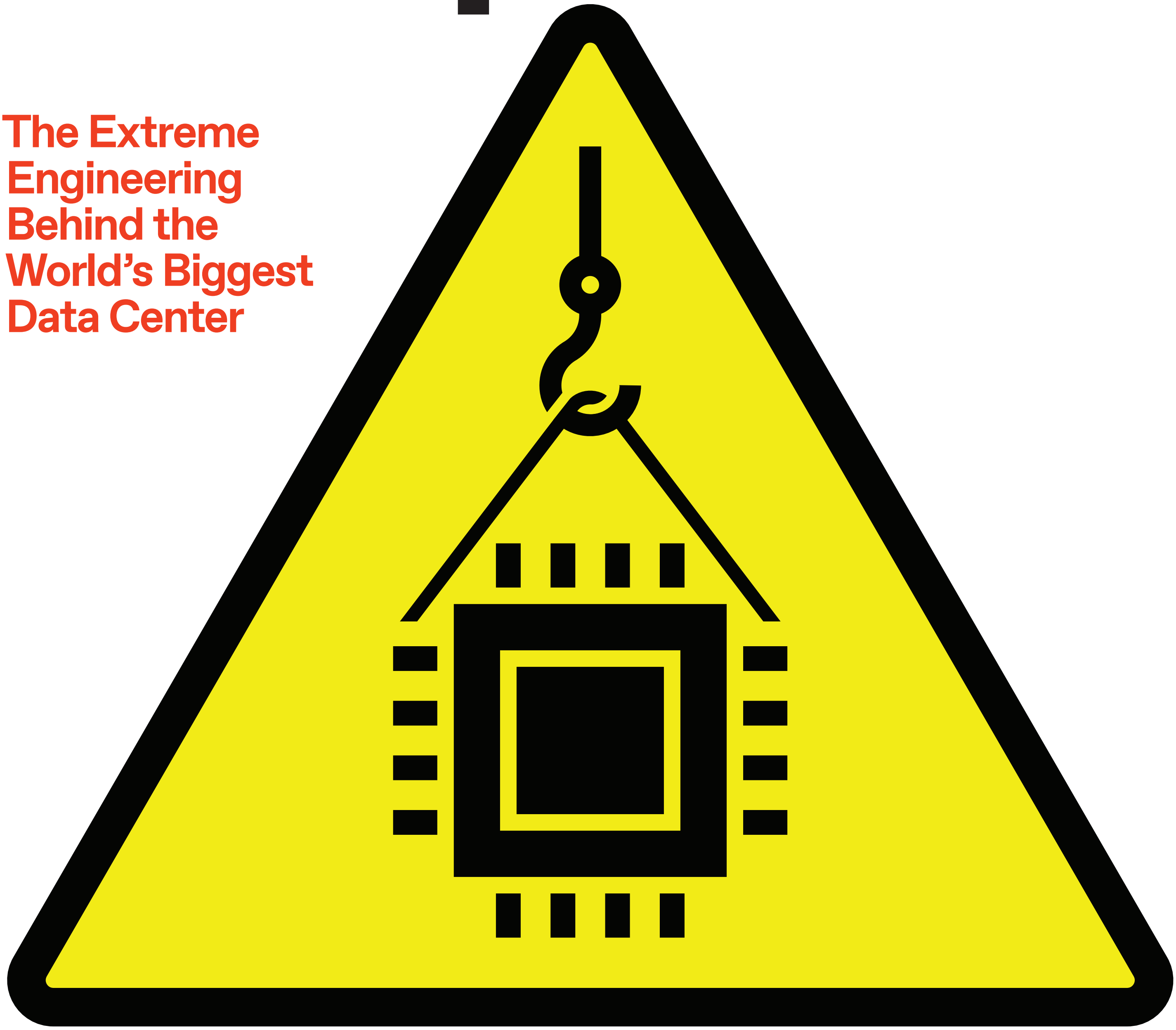
The Memory Chip Shortage, Explained › And why it won't be ending anytime soon
P. 28

The Promise of Smart Pills › They'll deliver drugs and take biopsies
P. 34

The Battlefield as Lab › War propels innovation in autonomous drones
P. 40

FOR THE TECHNOLOGY INSIDER
APRIL 2026

IEEE Spectrum



The Extreme Engineering Behind the World's Biggest Data Center

IEEE

LIVE
LIFE ON
YOUR
TERMS.
INSURANCE
THAT FITS
YOU.
Group Term Life Insurance
engineered for you.

1-800-493-IEEE (4333)

To learn more*, visit IEEEinsurance.com/10Year

IEEE
LIFE INSURANCE

*For information on features, costs, eligibility, renewability, limitations and exclusions.

Group Term Life Insurance is available only for residents of the U.S. (except territories), Puerto Rico and Canada (except Quebec). This is underwritten by New York Life Insurance Company, 51 Madison Ave., New York, NY 10010 on Policy Form GMR. AMBA does not act as broker with respect to Canadian residents and acts solely as an Administrator on behalf of New York Life.

Association Member Benefits Advisors, LLC.
In CA d/b/a Association Member Benefits & Insurance Agency
CA License #0I96562 • AR License #100114462
Program Administered by AMBA Administrators, Inc.

107581 (3/26) Copyright 2026 AMBA. All rights reserved

VOLUME 63 / ISSUE 4 APRIL 2026

Contents

40
Rise of the Autonomous Attack Drones
Combatants in Ukraine are pushing drone technology to the brink of autonomy.
By Tereza Pultarova

22
Giant Data Centers Change Everything
Engineers have to rethink power, cooling, concrete, and land.
By Matthew S. Smith

28
AI Is a Memory Hog
Its demand for DRAM threatens whole categories of electronics.
By Samuel K. Moore

34
When Pills Become Machines
Ingestible electronics turn capsules into medical devices.
By Reza Ghodssi, Justin Stine & Luke Beardslee

EDITOR'S NOTE 2
Ripples from AI's overconsumption are spreading far and wide.

NEWS 5
NASA's Lunar Struggles
Storing Data in Glass
Social Media for AI

THE DATA 14
They will know you by your face.

HANDS ON 16
A homebrew invention mixes spray paint like never before.

CAREERS 19
Is enrollment waning at U.S. Ph.D. programs?

5 QUESTIONS 21
Automakers design sounds for EVs.

PAST FORWARD 48
Deep Rover Dove Deep

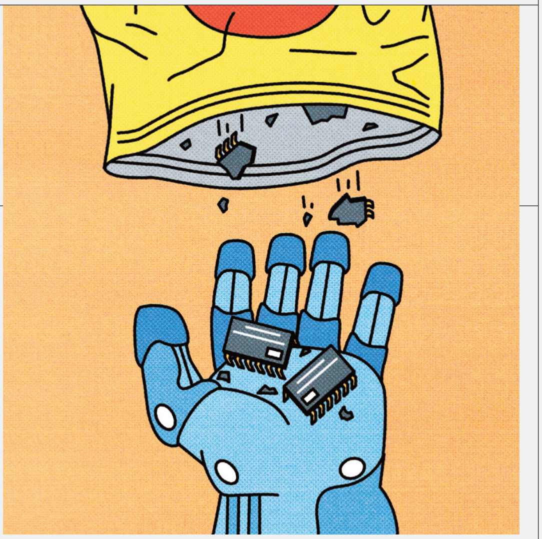
ON THE COVER:
Illustration by Optics Lab

FINBARR O'REILLY

BY HARRY GOLDSTEIN

AI Is Insatiable

And it's got the munchies for memory chips



While browsing our website a few weeks ago, I stumbled upon "How and When the Memory Chip Shortage Will End" by Senior Editor Samuel K. Moore. His analysis focuses on the current DRAM shortage caused by AI hyperscalers' ravenous appetite for memory, a major constraint on the speed at which large language models run. Moore provides a clear explanation of the shortage, particularly for high bandwidth memory (HBM).

As we and the rest of the tech media have documented, AI is a resource hog. AI electricity consumption could account for up to 12 percent of all U.S. power by 2028. Generative AI queries consumed 15 terawatt-hours in 2025 and are projected to consume 347 TWh by 2030. Water consumption for cooling AI data centers is predicted to double or even quadruple by 2028 compared to 2023.

But Moore's reporting shines a light on an obscure corner of the AI boom. HBM is a particular type of memory product tailor-made to serve AI processors. Makers of those processors, notably Nvidia and AMD, are demanding more and more memory for each of their chips, driven by the needs and wants of firms like Google, Microsoft, OpenAI, and Anthropic, which are underwriting an unprecedented build-out of data centers. And some of these facilities are colossal: You can read about the engineering challenges of building Meta's mind-boggling 5-gigawatt Hyperion site in Louisiana, in "What Will It Take to Build the World's Largest Data Center?" on page 22.

We realized that Moore's HBM story was both important and unique, and so we decided to include it in this issue, with some updates since the original published on 10 February. We paired it with a recent story by Contributing Editor Matthew S. Smith exploring how the memory-chip shortage is driving up the price of low-cost computers like the Raspberry Pi. The result is "AI Is a Memory Hog" [p. 28].

To be sure you don't miss any of Moore's analysis of this topic and to stay current on the entire spectrum of technology development, sign up for our weekly newsletter, Tech Alert, by scanning the QR code below.

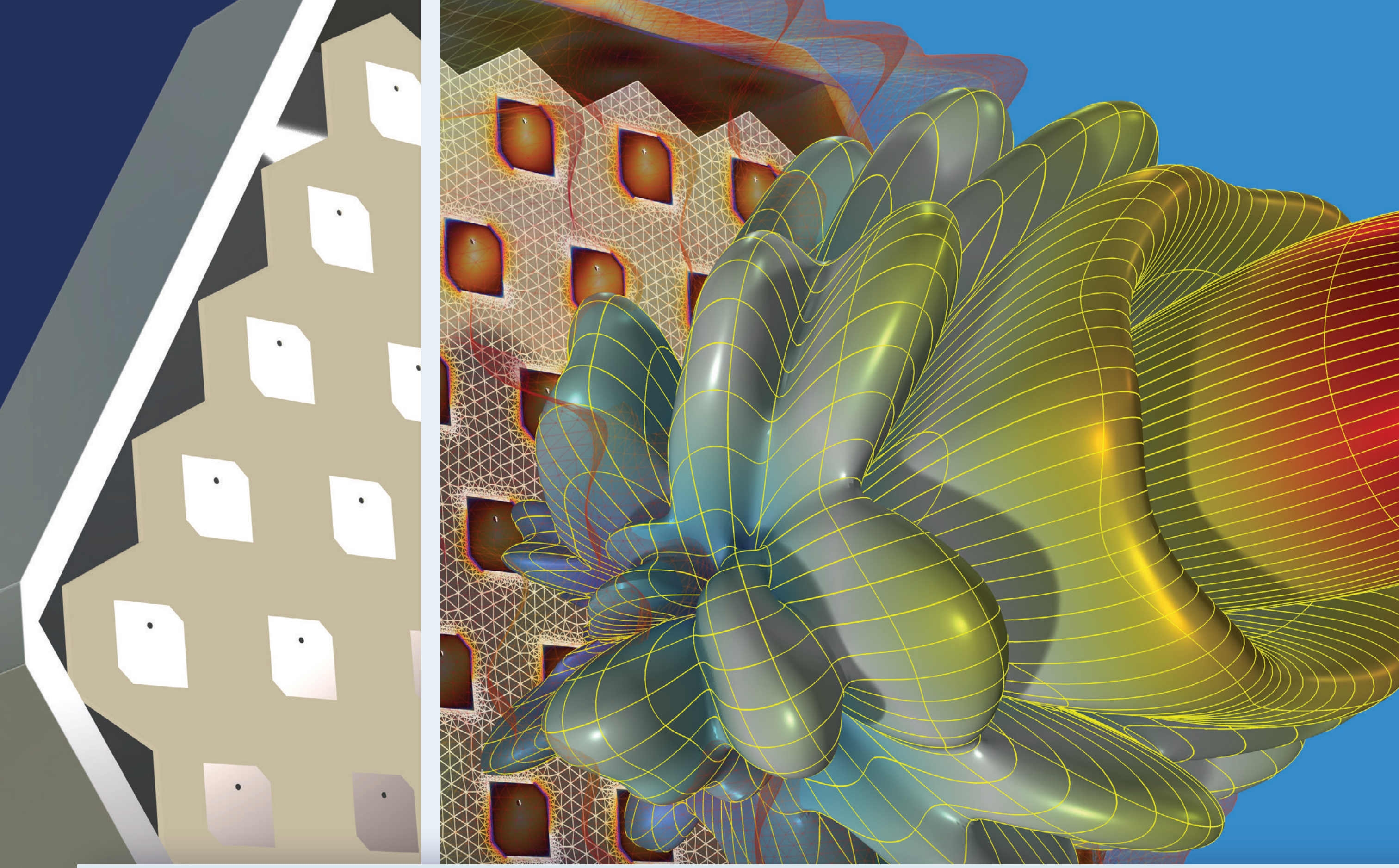
The big question now is, When will the shortage end? Price pressure caused by AI hyperscaler demand on all kinds of consumer electronics is being masked by stubborn inflation combined with a perpetually shifting tariff regime, at least here in the United States. So I asked Moore what indicators he's looking for that would signal an easing of the memory shortage.

"On the supply side, I'd say that if any of the big three HBM companies—Micron, Samsung, and SK Hynix—say that they are adjusting the schedule of the arrival of new production, that'd be an important signal," Moore told me. "On the demand side, it will be interesting to see how tech companies adapt up and down the supply chain. Data centers might steer toward hardware that sacrifices some performance for less memory. Startups developing all sorts of products might pivot toward creative redesigns that use less memory. Constraints like shortages can lead to interesting technology solutions, so I'm looking forward to covering those." ■

CORRECTION: Due to an editing error, a photograph in the March Hands On column was mislabeled. The individuals standing behind Jimi Hendrix are engineer Eddie Kramer [left] and studio manager Jim Marron, not Mitch Mitchell and Noel Redding.

PORTRAIT BY SERGIO ALBIAC; BOTTOM: FRED W. MCDARRAH/GETTY IMAGES

Illustration by George Wylesol



Take the Lead in RF Design

with COMSOL Multiphysics®

Multiphysics simulation is expanding the scope of RF analysis to higher frequencies and data rates. Accurate models of microwave, mmWave, and photonic designs are obtained by accounting for coupled physics effects, material property variation, and geometry deformation. Ultimately, this helps you more quickly see how a design will perform in the real world.

SCAN ME TO LEARN MORE

comsol.com/feature/rf-innovation

COMSOL

CONTRIBUTORS

● REZA GHODSSI

Ghodssi is the Herbert Rabin Distinguished Chair in Engineering at the University of Maryland, College Park, where he directs the MEMS Sensors and Actuators Laboratory. In his feature on ingestible electronics [p. 34], he describes efforts to build swallowable devices that can sense disease, deliver drugs, and collect tissue samples inside the gut. Ghodssi calls the gastrointestinal tract "one of the most challenging ecosystems" for miniature devices. His coauthors, Justin Stine and Luke Beardslee, are colleagues at the University of Maryland.

● NED POTTER

Potter, an *IEEE Spectrum* contributing editor, spent more than 25 years as a science and technology correspondent for ABC News and CBS News. This month, he writes about NASA's struggles to get its Artemis astronauts to the moon [p. 5]. He covered the very beginning of the effort, back in 2005, when NASA tried to save money by building its new moon rocket with used space shuttle components. As one engineer told him, "The best technology isn't the newest, it's the most proven."

● TEREZA PULTAROVA

Pultarova is a London-based journalist specializing in aerospace and defense technologies. A regular contributor to *IEEE Spectrum*, she has been following the emergence of autonomous technology in the drone-counterdrone struggle in Ukraine, a topic she revisits this month [p. 40]. "Having been born behind the Iron Curtain myself, I feel that I have an almost visceral understanding of what is at stake for the Ukrainians," she says. When not covering the war, she finds time to work on her first novel.

● MATTHEW S. SMITH

Smith is a technology journalist, author, and consultant based in Portland, Oregon. In this issue, Smith breaks down what it takes to bring a behemoth data center to life from an engineer's perspective [p. 22]. "I'd heard the scale of the AI data center build-out was immense, but when I took a closer look, it proved even larger than I thought. It includes millions of GPUs, tens of thousands of contractors, and sites so large they're going up alongside dedicated power plants that could handle a small city."

IEEE Spectrum

EDITOR IN CHIEF Harry Goldstein, h.goldstein@ieee.org
EXECUTIVE EDITOR Jean Kumagai, j.kumagai@ieee.org
MANAGING EDITOR Elizabeth A. Bretz, e.bretz@ieee.org
CREATIVE DIRECTOR Mark Montgomery, m.montgomery@ieee.org
DIRECTOR OF DIGITAL INNOVATION Erico Guizzo, e.guizzo@ieee.org
EDITORIAL DIRECTOR, CONTENT DEVELOPMENT
Glenn Zorpette, g.zorpette@ieee.org

FEATURES EDITOR Eliza Strickland, e.strickland@ieee.org

SPECIAL PROJECTS DIRECTOR Stephen Cass, cass.s@ieee.org

SENIOR EDITORS
Evan Ackerman, evan.ackerman@ieee.org
Lucas Laursen, l.laursen@ieee.org
Samuel K. Moore, s.k.moore@ieee.org

NEWS MANAGER Michael Koziol, m.koziol@ieee.org

SENIOR ASSOCIATE EDITORS
Margo Anderson, m.k.anderson@ieee.org
Emily Waltz, e.waltz@ieee.org

ASSOCIATE EDITORS
Dina Genkina, d.genkina@ieee.org
Willie D. Jones, w.jones@ieee.org

ASSISTANT EDITOR Gwendolyn Rak, g.rak@ieee.org

PRODUCT AND AUDIENCE DEVELOPMENT DIRECTOR
Tim Warder, t.warder@ieee.org
PRODUCT MANAGER Priyanka Shitole, p.shitole@ieee.org
AUDIENCE DEVELOPMENT MANAGER
Laura Bridgeman, l.bridgeman@ieee.org
COMMUNITY MANAGER
Kohava Mendelsohn, k.mendelsohn@ieee.org
DIGITAL PRODUCT DESIGNER Erik Vrielink, e.vrielink@ieee.org

ART & PRODUCTION
ART DIRECTOR Brandon Palacio, b.palacio@ieee.org
PHOTOGRAPHY DIRECTOR Randi Klett, randi.klett@ieee.org
ASSISTANT PHOTO EDITOR Nicole Millman, n.millman@ieee.org
PRINT PRODUCTION SPECIALIST
Sylvana Meneses, s.meneses@ieee.org
DIGITAL PRODUCTION MANAGER
Michael Spector, m.spector@ieee.org

SENIOR COPY EDITOR Joseph N. Levine, j.levine@ieee.org
COPY EDITOR Michele Kogon, m.kogon@ieee.org
RESEARCH COPY EDITOR Cheryl Della Pietra, c.dellapietra@ieee.org

ADMINISTRATOR Jose Zambrano, jose.zambrano@ieee.org

CONTRIBUTING EDITORS Rina Diane Caballar, Robert N. Charette, Charles Q. Choi, Tom Clynes, Peter Fairley, Edd Gent, W. Wayt Gibbs, Mark Harris, Allison Marsh, Julianne Pepitone, Ned Potter, Matthew S. Smith, Lawrence Ulrich

THE INSTITUTE
EDITOR IN CHIEF Kathy Pretz, k.pretz@ieee.org
ASSOCIATE EDITOR Joanna Goodrich, j.goodrich@ieee.org

SUBSCRIPTION SERVICES
IEEE Members member-services@ieee.org
Non-members +1 866 363 2304

DIRECTOR, PERIODICALS PRODUCTION SERVICES Peter Tuohy
ADVERTISING PRODUCTION MANAGER
Felicia Spagnoli, f.spagnoli@ieee.org

ADVERTISING PRODUCTION +1 732 562 6334

EDITORIAL ADVISORY BOARD, *IEEE SPECTRUM*
Harry Goldstein, *Chair*; Muhammad Alaa, Karanveer Anand, Ella M. Atkins, Sangyeun Cho, Hugh Durrant-Whyte, Benjamin Gross, Carlos Gutierrez, Lawrence O. Hall, Benjamin Kroposki, Somdeb Majumdar, Khaled Mokhtar, Ramune Nagisetty, Paul Nielsen, Ketra Schmidt, Gaurav Sharma, Fan Shi, Mini S. Thomas, Andrej Zdravkovic, Lidong Zhou

EDITORIAL ADVISORY BOARD, *THE INSTITUTE*
Kathy Pretz, *Chair*; Stamatis Dragoumanos, Madeleine Glick, Harry Goldstein, Lawrence O. Hall, Francesca Iacopi, Sai Prashanth Mallellu, Shashi Raj Pandey, John Purvis, Chenyang Xu

EDITORIAL CORRESPONDENCE
IEEE Spectrum, 3 Park Ave., 17th Floor,
New York, NY 10016-5997 Tel: +1 212 419 7555

ADVERTISING INQUIRIES Wiley
Joerg Wuellner +49 152 2258 4972, jwuellner@wiley.com
Chris Darch +1 206 867 4165, cdarch@wiley.com

REPRINT SALES +1 212 221 9595, ext. 319

REPRINT PERMISSION/ LIBRARIES Articles may be photocopied for private use of patrons. A per-copy fee must be paid to the Copyright Clearance Center, 29 Congress St., Salem, MA 01970. For other copying or republication, contact Managing Editor, *IEEE Spectrum.*

COPYRIGHTS AND TRADEMARKS *IEEE Spectrum* is a registered trademark owned by The Institute of Electrical and Electronics Engineers Inc. Responsibility for the substance of articles rests upon the authors, not IEEE, its organizational units, or its members. Articles do not represent official positions of IEEE. Readers may post comments online; comments may be excerpted for publication. IEEE reserves the right to reject any advertising.

IEEE

IEEE BOARD OF DIRECTORS
PRESIDENT&CEO Mary Ellen Randall, president@ieee.org
+1 732 562 3928 Fax: +1 732 981 9515
PRESIDENT-ELECT Jill I. Gostin
PAST PRESIDENT Kathleen A. Kramer
TREASURER Gerardo Barbosa
SECRETARY Jenifer Castillo Rodriguez

VICE PRESIDENTS
Timothy P. Kurzweg, *Educational Activities;* Andrew D. Lowery, *Member & Geographic Activities;* W. Clem Karl, *Publication Services & Products;* F.D. Tan, *Technical Activities;* Gary R. Hoffman, *President, Standards Association;* Barry C. Tilton, *President, IEEE-USA*

DIVISION DIRECTORS
Fernando J. Guarin (I); H. Alan Mantooth (II); Sarah Kate Wilson (III); Charles M. Jackson (IV); Leila De Floriani (V); Russell D. Meier (VI); Noel M. Schulz (VII); Cecilia Metra (VIII); Cristina Olaverri-Monreal (IX); Ravinder S. Dahiya (X)

REGION DIRECTORS
Charles P. Rubenstein (1); Felicia Reinhart (2); Sonya L. Dillard (3); Constance A. Kelly (4); Christopher B. Sanderson (5); Joseph C. Wei (6); Wahab Almuhtadi (7); Dusanka Boskovic (8); Jose-Ignacio Castillo-Velazquez (9); Takako Hashimoto (10)

IEEE STAFF
EXECUTIVE DIRECTOR & COO Sophia A. Muirhead
+1 732 562 5400, s.muirhead@ieee.org
CHIEF FINANCIAL OFFICER Kelly Armstrong
+1 732 562 6011, k.armstrong@ieee.org
MANAGING DIRECTOR,TECHNICAL ACTIVITIES Mojdeh Bahar
+1 732 562 3850, m.bahar@ieee.org
GENERAL COUNSEL & CHIEF COMPLIANCE OFFICER
Ahsaki Benion +1 212 705 8975, a.benion@ieee.org
MANAGING DIRECTOR, PUBLICATIONS Steven Heffner
+1 212 705 8958, s.heffner@ieee.org
CHIEF GOVERNANCE OFFICER Donna Hourican
+1 732 562 6330, d.hourican@ieee.org
MANAGING DIRECTOR, CONFERENCES, EVENTS AND EXPERIENCES Marie Hunter
+ 1 732 465 5889, m.m.hunter@ieee.org
MANAGING DIRECTOR, MEMBER & GEOGRAPHIC ACTIVITIES
Cecelia Jankowski +1 732 562 5504, c.jankowski@ieee.org
MANAGING DIRECTOR, EDUCATIONAL ACTIVITIES
Jamie Moesch +1 732 562 5514, j.moesch@ieee.org
CHIEF MARKETING OFFICER Jayne O'Brien
+1 732 981 3433, jayne.obrien@ieee.org
MANAGING DIRECTOR, IEEE STANDARDS Alpesh Shah
+1 732 465 6467, alpesh.shah@ieee.org
CHIEF INFORMATION AND DIGITAL OFFICER Jeff Strohschein
+1 732 562 5332, j.strohschein@ieee.org
CHIEF OF STAFF Allison Taylor
+1 732 562 6068, allison.taylor@ieee.org
CHIEF HUMAN RESOURCES OFFICER Cheri N. Collins Wideman
+1 732 562 5335, c.wideman@ieee.org
MANAGING DIRECTOR, IEEE-USA Russell T. Harrison
+1 202 530 8326, r.t.harrison@ieee.org

IEEE PUBLICATION SERVICES & PRODUCTS BOARD
W. Clem Karl, *Chair*; John T. Barr, Nazanin Bassiri-Gharb, Sergio Benedetto, Eddie Custovic, Stephen Dukes, Alessandro Foi, Lawrence O. Hall, Charles Hansen, Warunika Hippola, James Irvine, Michael Jensen, Athanasios Kakarountas, James Keller, Imran Mehdi, Khaled Mokhtar, Paolo Montuschi, Zorana Popovic, Mehrdad Saif, Gianluca Setti, Gaurav Sharma, Sarah Spurgeon, Jaideep Vaidya, Maria Elena Valcher, Patrick Wheeler, Steve Yurkovich

IEEE OPERATIONS CENTER
445 Hoes Lane, Box 1331, Piscataway, NJ 08854-1331 U.S.A.
Tel: +1 732 981 0060 Fax: +1 732 981 1721

IEEE SPECTRUM (ISSN 0018-9235) is published monthly by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. © 2026 by The Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, New York, NY 10016-5997, U.S.A. Volume No. 63, Issue No. 4. The editorial content of *IEEE Spectrum* magazine does not represent official positions of the IEEE or its organizational units. Canadian Post International Publications Mail (Canadian Distribution) Sales Agreement No. 40013087. Return undeliverable Canadian addresses to: Circulation Department, *IEEE Spectrum*, Box 1051, Fort Erie, ON L2A 6C7. Cable address: ITRIPLEE. Fax: +1 212 419 7570. INTERNET: spectrum@ieee.org. ANNUAL SUBSCRIPTIONS: United States and Canada: $74.95. Elsewhere: $99.95. POSTMASTER: Please send address changes to *IEEE Spectrum*, ℅ Coding Department, IEEE Service Center, 445 Hoes Lane, Box 1331, Piscataway, NJ 08855. Periodicals postage paid at New York, NY, and additional mailing offices. Canadian GST #125634188. Printed at 3401 N. Heartland Dr., Liberty, MO 64068, U.S.A. Printed at Spenta Multimedia Pvt. Ltd., Plot 15, 16 & 21/1, Village Chikhloli, Morivali, MIDC, Ambernath (West), Dist. Thane. *IEEE Spectrum* circulation is audited by BPA Worldwide. *IEEE Spectrum* is a member of the Association of Business Information & Media Companies, the Association of Magazine Media, and Association Media & Publishing. IEEE prohibits discrimination, harassment, and bullying. For more information, visit https://www.ieee.org/about/corporate/governance/p9-26.html.

CONTRIBUTORS

● REZA GHODSSI

Ghodssi is the Herbert Rabin Distinguished Chair in Engineering at the University of Maryland, College Park, where he directs the MEMS Sensors and Actuators Laboratory. In his feature on ingestible electronics [p. 34], he describes efforts to build swallowable devices that can sense disease, deliver drugs, and collect tissue samples inside the gut. Ghodssi calls the gastrointestinal tract “one of the most challenging ecosystems” for miniature devices. His coauthors, Justin Stine and Luke Beardslee, are colleagues at the University of Maryland.

● NED POTTER

Potter, an *IEEE Spectrum* contributing editor, spent more than 25 years as a science and technology correspondent for ABC News and CBS News. This month, he writes about NASA’s struggles to get its Artemis astronauts to the moon [p. 5]. He covered the very beginning of the effort, back in 2005, when NASA tried to save money by building its new moon rocket with used space shuttle components. As one engineer told him, “The best technology isn’t the newest, it’s the most proven.”

● TEREZA PULTAROVA

Pultarova is a London-based journalist specializing in aerospace and defense technologies. A regular contributor to *IEEE Spectrum*, she has been following the emergence of autonomous technology in the drone-counterdrone struggle in Ukraine, a topic she revisits this month [p. 40]. “Having been born behind the Iron Curtain myself, I feel that I have an almost visceral understanding of what is at stake for the Ukrainians,” she says. When not covering the war, she finds time to work on her first novel.

● MATTHEW S. SMITH

Smith is a technology journalist, author, and consultant based in Portland, Oregon. In this issue, Smith breaks down what it takes to bring a behemoth data center to life from an engineer’s perspective [p. 22]. “I’d heard the scale of the AI data center build-out was immense, but when I took a closer look, it proved even larger than I thought. It includes millions of GPUs, tens of thousands of contractors, and sites so large they’re going up alongside dedicated power plants that could handle a small city.”

IEEE Spectrum

EDITOR IN CHIEF Harry Goldstein, h.goldstein@ieee.org
EXECUTIVE EDITOR Jean Kumagai, j.kumagai@ieee.org
MANAGING EDITOR Elizabeth A. Bretz, e.bretz@ieee.org
CREATIVE DIRECTOR Mark Montgomery, m.montgomery@ieee.org
DIRECTOR OF DIGITAL INNOVATION Erico Guizzo, e.guizzo@ieee.org
EDITORIAL DIRECTOR, CONTENT DEVELOPMENT
Glenn Zorpette, g.zorpette@ieee.org

FEATURES EDITOR Eliza Strickland, e.strickland@ieee.org

SPECIAL PROJECTS DIRECTOR Stephen Cass, cass.s@ieee.org

SENIOR EDITORS
Evan Ackerman, evan.ackerman@ieee.org
Lucas Laursen, l.laursen@ieee.org
Samuel K. Moore, s.k.moore@ieee.org

NEWS MANAGER Michael Koziol, m.koziol@ieee.org

SENIOR ASSOCIATE EDITORS
Margo Anderson, m.k.anderson@ieee.org
Emily Waltz, e.waltz@ieee.org

ASSOCIATE EDITORS
Dina Genkina, d.genkina@ieee.org
Willie D. Jones, w.jones@ieee.org

ASSISTANT EDITOR Gwendolyn Rak, g.rak@ieee.org

PRODUCT AND AUDIENCE DEVELOPMENT DIRECTOR
Tim Warder, t.warder@ieee.org
PRODUCT MANAGER Priyanka Shitole, p.shitole@ieee.org
AUDIENCE DEVELOPMENT MANAGER
Laura Bridgeman, l.bridgeman@ieee.org
COMMUNITY MANAGER
Kohava Mendelsohn, k.mendelsohn@ieee.org
DIGITAL PRODUCT DESIGNER Erik Vrielink, e.vrielink@ieee.org

ART & PRODUCTION
ART DIRECTOR Brandon Palacio, b.palacio@ieee.org
PHOTOGRAPHY DIRECTOR Randi Klett, randi.klett@ieee.org
ASSISTANT PHOTO EDITOR Nicole Millman, n.millman@ieee.org
PRINT PRODUCTION SPECIALIST
Sylvana Meneses, s.meneses@ieee.org
DIGITAL PRODUCTION MANAGER
Michael Spector, m.spector@ieee.org

SENIOR COPY EDITOR Joseph N. Levine, j.levine@ieee.org
COPY EDITOR Michele Kogon, m.kogon@ieee.org
RESEARCH COPY EDITOR Cheryl Della Pietra, c.dellapietra@ieee.org

ADMINISTRATOR Jose Zambrano, jose.zambrano@ieee.org

CONTRIBUTING EDITORS Rina Diane Caballar, Robert N. Charette, Charles Q. Choi, Tom Clynes, Peter Fairley, Edd Gent, W. Wayt Gibbs, Mark Harris, Allison Marsh, Julianne Pepitone, Ned Potter, Matthew S. Smith, Lawrence Ulrich

THE INSTITUTE
EDITOR IN CHIEF Kathy Pretz, k.pretz@ieee.org
ASSOCIATE EDITOR Joanna Goodrich, j.goodrich@ieee.org

SUBSCRIPTION SERVICES
IEEE Members member-services@ieee.org
Non-members +1 866 363 2304

DIRECTOR, PERIODICALS PRODUCTION SERVICES Peter Tuohy
ADVERTISING PRODUCTION MANAGER
Felicia Spagnoli, f.spagnoli@ieee.org

ADVERTISING PRODUCTION +1 732 562 6334

EDITORIAL ADVISORY BOARD, *IEEE SPECTRUM*
Harry Goldstein, *Chair*; Muhammad Alaa, Karanveer Anand, Ella M. Atkins, Sangyeun Cho, Hugh Durrant-Whyte, Benjamin Gross, Carlos Gutierrez, Lawrence O. Hall, Benjamin Kroposki, Somdeb Majumdar, Khaled Mokhtar, Ramune Nagisetty, Paul Nielsen, Ketra Schmidt, Gaurav Sharma, Fan Shi, Mini S. Thomas, Andrej Zdravkovic, Lidong Zhou

EDITORIAL ADVISORY BOARD, *THE INSTITUTE*
Kathy Pretz, *Chair*; Stamatis Dragoumanos, Madeleine Glick, Harry Goldstein, Lawrence O. Hall, Francesca Iacopi, Sai Prashanth Mallellu, Shashi Raj Pandey, John Purvis, Chenyang Xu

EDITORIAL CORRESPONDENCE
IEEE Spectrum, 3 Park Ave., 17th Floor,
New York, NY 10016-5997 Tel: +1 212 419 7555

ADVERTISING INQUIRIES Wiley
Joerg Wuellner +49 152 2258 4972, jwuellner@wiley.com
Chris Darch +1 206 867 4165, cdarch@wiley.com

REPRINT SALES +1 212 221 9595, ext. 319

REPRINT PERMISSION/ LIBRARIES Articles may be photocopied for private use of patrons. A per-copy fee must be paid to the Copyright Clearance Center, 29 Congress St., Salem, MA 01970. For other copying or republication, contact Managing Editor, *IEEE Spectrum.*

COPYRIGHTS AND TRADEMARKS *IEEE Spectrum* is a registered trademark owned by The Institute of Electrical and Electronics Engineers Inc. Responsibility for the substance of articles rests upon the authors, not IEEE, its organizational units, or its members. Articles do not represent official positions of IEEE. Readers may post comments online; comments may be excerpted for publication. IEEE reserves the right to reject any advertising.

IEEE

IEEE BOARD OF DIRECTORS
PRESIDENT&CEO Mary Ellen Randall, president@ieee.org
+1 732 562 3928 Fax: +1 732 981 9515
PRESIDENT-ELECT Jill I. Gostin
PAST PRESIDENT Kathleen A. Kramer
TREASURER Gerardo Barbosa
SECRETARY Jenifer Castillo Rodriguez

VICE PRESIDENTS
Timothy P. Kurzweg, *Educational Activities;* Andrew D. Lowery, *Member & Geographic Activities;* W. Clem Karl, *Publication Services & Products;* F.D. Tan, *Technical Activities;* Gary R. Hoffman, *President, Standards Association;* Barry C. Tilton, *President, IEEE-USA*

DIVISION DIRECTORS
Fernando J. Guarin (I); H. Alan Mantooth (II); Sarah Kate Wilson (III); Charles M. Jackson (IV); Leila De Floriani (V); Russell D. Meier (VI); Noel M. Schulz (VII); Cecilia Metra (VIII); Cristina Olaverri-Monreal (IX); Ravinder S. Dahiya (X)

REGION DIRECTORS
Charles P. Rubenstein (1); Felicia Reinhart (2); Sonya L. Dillard (3); Constance A. Kelly (4); Christopher B. Sanderson (5); Joseph C. Wei (6); Wahab Almuhtadi (7); Dusanka Boskovic (8); Jose-Ignacio Castillo-Velazquez (9); Takako Hashimoto (10)

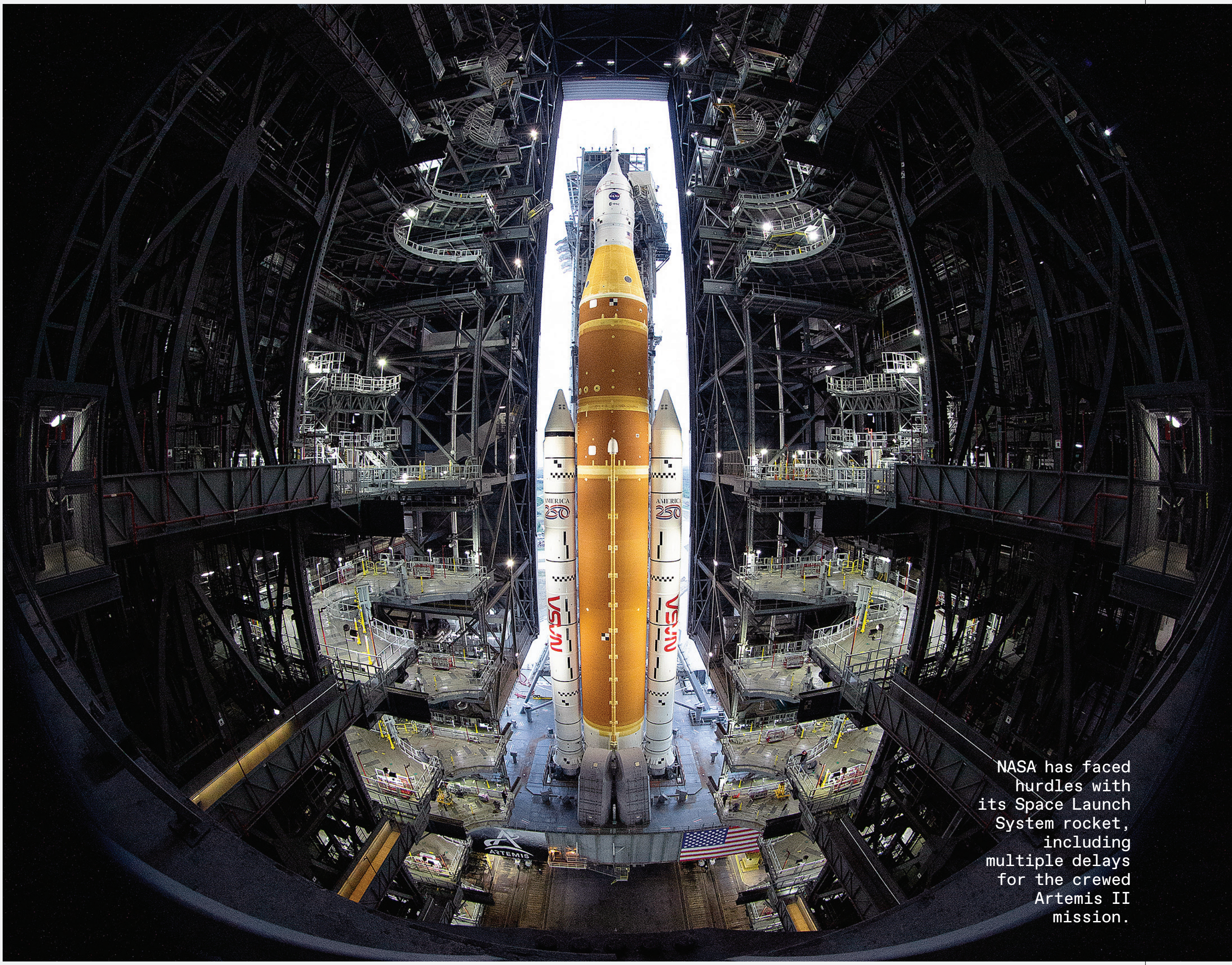
IEEE STAFF
EXECUTIVE DIRECTOR & COO Sophia A. Muirhead
+1 732 562 5400, s.muirhead@ieee.org
CHIEF FINANCIAL OFFICER Kelly Armstrong
+1 732 562 6011, k.armstrong@ieee.org
MANAGING DIRECTOR,TECHNICAL ACTIVITIES Mojdeh Bahar
+1 732 562 3850, m.bahar@ieee.org
GENERAL COUNSEL & CHIEF COMPLIANCE OFFICER
Ahsaki Benion +1 212 705 8975, a.benion@ieee.org
MANAGING DIRECTOR, PUBLICATIONS Steven Heffner
+1 212 705 8958, s.heffner@ieee.org
CHIEF GOVERNANCE OFFICER Donna Hourican
+1 732 562 6330, d.hourican@ieee.org
MANAGING DIRECTOR, CONFERENCES, EVENTS AND EXPERIENCES Marie Hunter
+ 1 732 465 5889, m.m.hunter@ieee.org
MANAGING DIRECTOR, MEMBER & GEOGRAPHIC ACTIVITIES
Cecelia Jankowski +1 732 562 5504, c.jankowski@ieee.org
MANAGING DIRECTOR, EDUCATIONAL ACTIVITIES
Jamie Moesch +1 732 562 5514, j.moesch@ieee.org
CHIEF MARKETING OFFICER Jayne O’Brien
+1 732 981 3433, jayne.obrien@ieee.org
MANAGING DIRECTOR, IEEE STANDARDS Alpesh Shah
+1 732 465 6467, alpesh.shah@ieee.org
CHIEF INFORMATION AND DIGITAL OFFICER Jeff Strohschein
+1 732 562 5332, j.strohschein@ieee.org
CHIEF OF STAFF Allison Taylor
+1 732 562 6068, allison.taylor@ieee.org
CHIEF HUMAN RESOURCES OFFICER Cheri N. Collins Wideman
+1 732 562 5335, c.wideman@ieee.org
MANAGING DIRECTOR, IEEE-USA Russell T. Harrison
+1 202 530 8326, r.t.harrison@ieee.org

IEEE PUBLICATION SERVICES & PRODUCTS BOARD
W. Clem Karl, *Chair*; John T. Barr, Nazanin Bassiri-Gharb, Sergio Benedetto, Eddie Custovic, Stephen Dukes, Alessandro Foi, Lawrence O. Hall, Charles Hansen, Warunika Hippola, James Irvine, Michael Jensen, Athanasios Kakarountas, James Keller, Imran Mehdi, Khaled Mokhtar, Paolo Montuschi, Zorana Popovic, Mehrdad Saif, Gianluca Setti, Gaurav Sharma, Sarah Spurgeon, Jaideep Vaidya, Maria Elena Valcher, Patrick Wheeler, Steve Yurkovich

IEEE OPERATIONS CENTER
445 Hoes Lane, Box 1331, Piscataway, NJ 08854-1331 U.S.A.
Tel: +1 732 981 0060 Fax: +1 732 981 1721

IEEE SPECTRUM (ISSN 0018-9235) is published monthly by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. © 2026 by The Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, New York, NY 10016-5997, U.S.A. Volume No. 63, Issue No. 4. The editorial content of *IEEE Spectrum* magazine does not represent official positions of the IEEE or its organizational units. Canadian Post International Publications Mail (Canadian Distribution) Sales Agreement No. 40013087. Return undeliverable Canadian addresses to: Circulation Department, *IEEE Spectrum*, Box 1051, Fort Erie, ON L2A 6C7. Cable address: ITRIPLEE. Fax: +1 212 419 7570. INTERNET: spectrum@ieee.org. ANNUAL SUBSCRIPTIONS: United States and Canada: $74.95. Elsewhere: $99.95. POSTMASTER: Please send address changes to *IEEE Spectrum*, % Coding Department, IEEE Service Center, 445 Hoes Lane, Box 1331, Piscataway, NJ 08855. Periodicals postage paid at New York, NY, and additional mailing offices. Canadian GST #125634188. Printed at 3401 N. Heartland Dr., Liberty, MO 64068, U.S.A. Printed at Spenta Multimedia Pvt. Ltd., Plot 15, 16 & 21/1, Village Chikhloli, Morivali, MIDC, Ambernath (West), Dist. Thane. *IEEE Spectrum* circulation is audited by BPA Worldwide. *IEEE Spectrum* is a member of the Association of Business Information & Media Companies, the Association of Magazine Media, and Association Media & Publishing. IEEE prohibits discrimination, harassment, and bullying. For more information, visit https://www.ieee.org/about/corporate/governance/p9-26.html.

News



NASA has faced hurdles with its Space Launch System rocket, including multiple delays for the crewed Artemis II mission.

JOEL KOWSKY/NASA

AEROSPACE

NASA's Lunar Return Struggles › Rescheduled launches and lander conundrums dog the project

BY NED POTTER

Artemis is NASA's bid to reclaim its heritage, to resume the business of human exploration, to take astronauts to the moon and beyond, and to win what's been billed by some U.S. politicians as a "new space race," with China. Artemis II is planned as the project's first circumlunar test mission with a crew. If it succeeds, and if NASA can deliver on a December executive order from the Trump administration, Artemis astronauts will land near the moon's south pole by 2028 and start building a lunar base by 2030, steps to ensure, as the order says, "American space superiority."

But Artemis II missed launch windows in February and March, the latest in a long string of delays. It's faced some of the same problems engineers thought they'd solved in the three-year postmortem for Artemis I, a flight without astronauts in 2022. Unless things turn around, say some former NASA managers, any race with China has already been lost.

> **"We can only control what we are doing. Of those efforts, I am forced to say that mediocrity would be an improvement."**
>
> —MICHAEL GRIFFIN, FORMER NASA ADMINISTRATOR

"We cannot control what China is doing," said Michael Griffin in congressional testimony in December. He was NASA administrator from 2005 to 2009, when the agency began assembling the hardware for what is now Artemis. "We can only control what we are doing. Of those efforts, I am forced to say that mediocrity would be an improvement."

Other NASA veterans have expressed the same worry. So far this year, Artemis has mostly been hobbled by fueling problems with its much-delayed Space Launch System (SLS) rocket. But, they say, it's also struggling with its Orion crew capsule, and—perhaps most critically—two competing, unproven lunar landers. They fault the program's history of convoluted, meandering decision-making. These former NASA officials say Artemis needs better organization and perhaps—even at this late date—a simplified landing ship.

Enter Jared Isaacman, NASA's new chief. In February, he rearranged the Artemis schedule, trying to make flights simpler and more frequent, adding components step-by-step instead of all at once. Artemis III, instead of being the first landing attempt, becomes a 2027 tryout of the prototype landing ships in low Earth orbit. The first landing missions would then be Artemis IV and V, hopefully in 2028. In addition, several upgrades to Artemis equipment have been canceled to save time and money. "SLS is a very impressive vehicle," Isaacman said when he announced the plan. "We don't want to turn every one of them into a work of art."

To NASA watchers, this comes none too soon. The Chinese space program claims it's on track to a lunar landing by 2030—uncomfortably close for the United States if its missions keep getting delayed. In Western eyes, China does not have superior technology or resources, but it's been better at setting long-term goals and sticking to them.

"If we don't do this, it's not just 'Oh, that's too bad, China got to the moon first,' " says Lisa Porter, who worked with Griffin at NASA and has cofounded a technology consulting firm, LogiQ, with him. Coming in second, she says, means "China gets to establish the standards for the future."

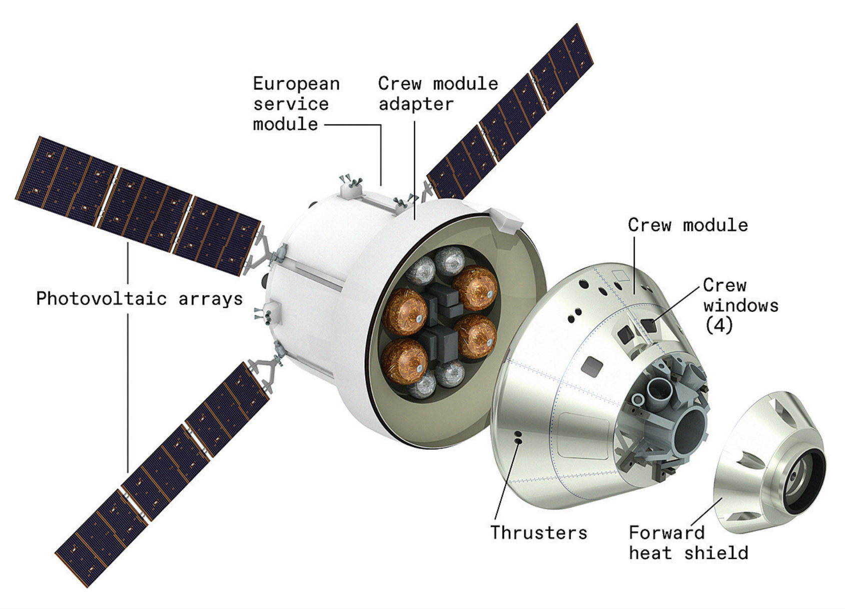
Artemis rose from the ashes of the space shuttle *Columbia* disaster. After the shuttle disintegrated upon reentry in 2003, President George W. Bush dispatched Griffin to lay out a new program called Constellation, which would put astronauts on the moon by 2018 or 2020.

That plan lasted only until the Bush administration was succeeded by Barack Obama's—which canceled Constellation as its costs spiraled. The Trump administration reversed Obama's reversal in 2017 and told NASA to get to the moon by 2024.

"There's a regrettable history in this country," says Griffin now, "of, whenever possible, a new administration wants to discontinue what was being done before and do something new. Sometimes that's good and sometimes it's not."

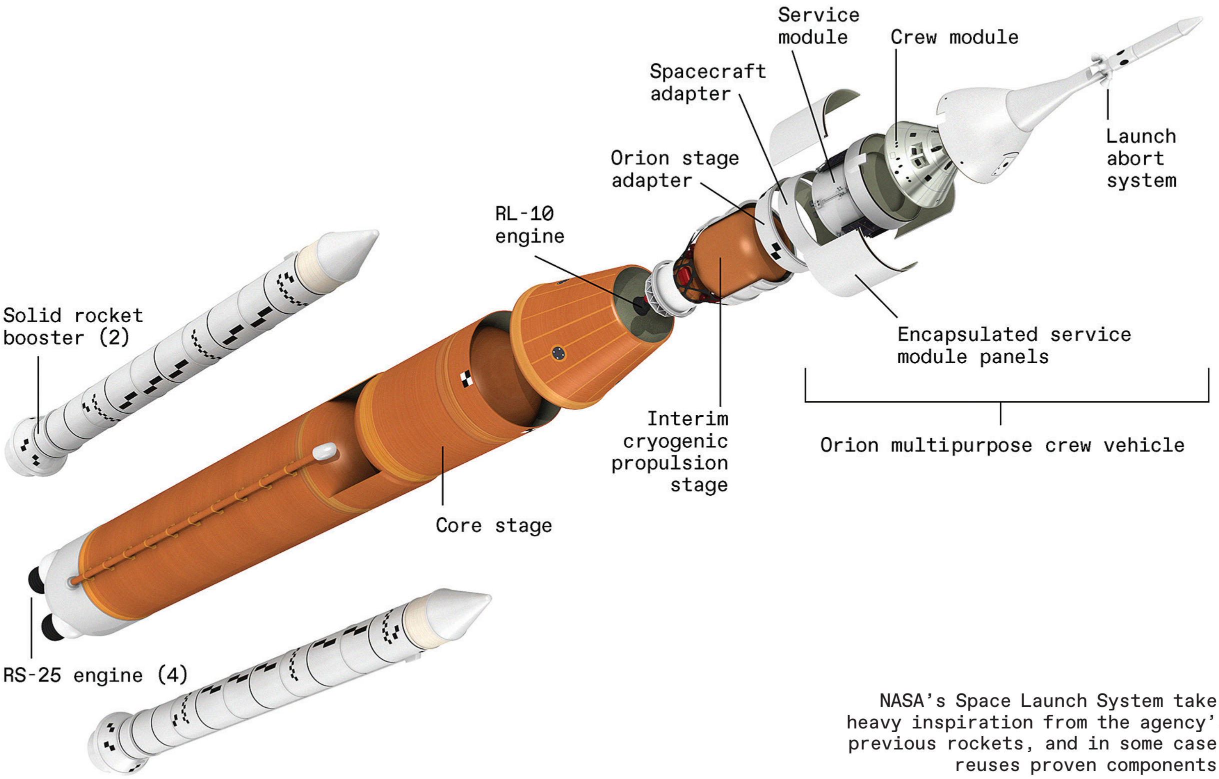
There has been a sea change, in the meantime, in how the U.S. space effort does business. No longer does NASA try to run every program itself; where possible, it lets industry manage them, with the government as a "customer."

Proponents say it works great: Companies, competing for NASA's business, have incentive to move faster and be more creative. Costs drop, sometimes so dramatically that NASA can afford to hire two companies for a job in case one runs into trouble. But in the case



Artemis astronauts will travel aboard the Orion spacecraft during lunar trips.

Illustrations by John MacNeill



NASA's Space Launch System takes heavy inspiration from the agency's previous rockets, and in some cases reuses proven components.

of Artemis's lunar landers, so far it's been messy.

In 2021, NASA awarded seed money to SpaceX, Elon Musk's company, to adapt its giant Starship spacecraft as a landing ship. Two years later, it made a second award, to Jeff Bezos's Blue Origin, to build a lander for later Artemis flights. The idea, in each case, is that the Orion crew capsule would rendezvous with the landing ship in lunar orbit for a descent to the surface. They were two aggressive companies, owned by two of the world's wealthiest technologists.

But that aggressiveness is why all those NASA veterans were sounding the alarm. Neither Starship nor Blue Origin's Blue Moon landers have yet proved themselves ready to carry astronauts. And, Griffin and Porter say, they're going to have a very difficult time even getting close.

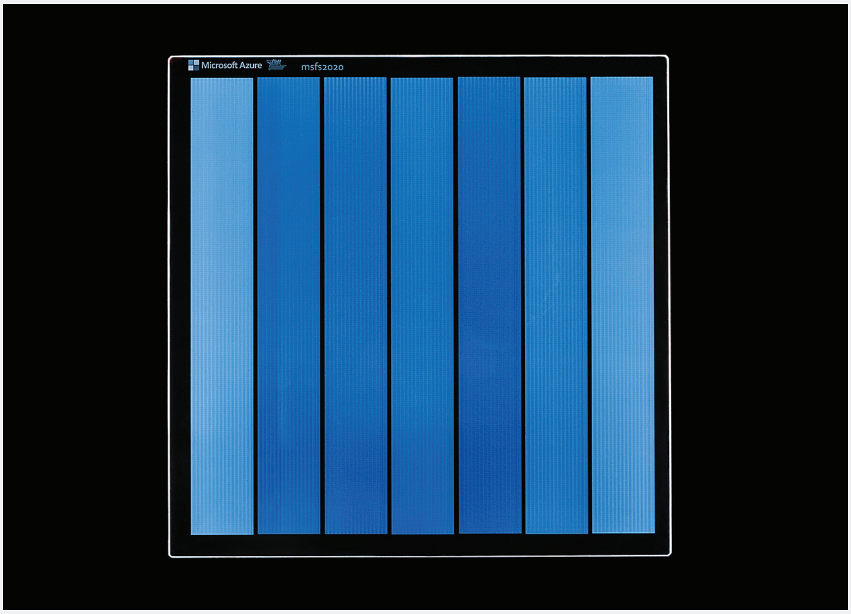
The biggest showstopper for the landers from SpaceX and Blue Origin is in the choice of propellants. For 60 years, U.S. spacecraft have maneuvered in space mostly with fuels called hypergolics, a group that includes hydrazine and nitrogen tetroxide, which ignite on contact with each other to provide thrust. They're highly toxic, but very reliable. SpaceX and Blue Origin both favor cryogenic fuels, like liquid hydrogen or liquid methane.

That makes great sense for the long term—cryogenic fuels have a much higher specific impulse (the amount of thrust per unit of propellant) than hypergolics, and someday they may even be made from water ice near the moon's south pole—but to be kept liquid in a rocket's fuel tank, they have to be hundreds of degrees below zero. Not an easy problem when the sun routinely heats orbiting spacecraft above 120 °C. To fill a moon-bound Starship with liquid methane, SpaceX proposed launching it into low Earth orbit, then sending a succession of other Starships to dock with it and tank it up. It would be an uphill battle, because the fuel just delivered would keep boiling off in the sun's heat.

How many Starship tankers would it take to fill one Starship lander in orbit? Four? Ten? Twenty? Nobody has ever tried large-scale refueling in orbit—much less with multiple launches of the largest rocket ever built. "It's an odd choice from a program manager's point of view to put an unproven technology in series with an operational mission," says Griffin.

Under pressure from NASA, SpaceX and Blue Origin have both promised to speed the work along on their landers. SpaceX said online, "We've shared and are formally assessing a simplified mission architecture and concept of operations that we believe will result in a faster return to the moon." And Blue Origin said it is stopping its short suborbital flights for space tourists so that it can "shift resources to further accelerate development of the company's human lunar capabilities."

This is exactly the kind of push NASA's Isaacman said he wanted. "I think the best thing for SpaceX is a Blue Origin right on their heels and vice versa," he said at his Senate confirmation hearing on 3 December 2025. And he took it a step further: "I think competition among world powers is actually a really good thing, just as long as we don't lose." ■



Microsoft researchers wrote the map data for Microsoft Flight Simulator onto this 12-centimeter-square piece of glass using a femtosecond laser.

MICROSOFT RESEARCH

COMPUTING

Microsoft Boosts Density of Forever Storage in Glass

› A single 12-centimeter-square chip could store 2 million books

BY CHARLES Q. CHOI

The growing amount of data in the world has driven research toward radically new ways of storing data, in part because all mainstream electronic data-storage techniques break down over time. For example, most digital archives are currently saved on magnetic tape, which generally lasts about 30 years.

But there's one possibility that could exist functionally forever: glass. The material is resistant to moisture, temperature fluctuations, and electromagnetic interference. At normal storage conditions, "glass and the data structures within it are essentially unchanged for far longer than human history," says Richard Black, the research director for Microsoft's Project Silica.

The project is Microsoft's effort to overcome existing challenges around data density, writing throughput, and energy efficiency that plague glass as a storage medium. Microsoft scientists are working to overcome those deficiencies by using femtosecond lasers, which fire high-power pulses of photons for just quadrillionths of a second. The researchers focused a laser beam on a point within a glass chip to modify its optical properties and encode data as voxels, the 3D equivalent of pixels. The researchers published their findings online on 18 February in *Nature*.

The voxels were about 0.5 micrometers across and separated from one another by about 6 micrometers. The data could be read out using an automated microscope with a camera. Machine learning models helped account for random errors and optical interference between voxels.

Project Silica can store up to 1.59 gigabits per cubic millimeter. This amounts to 4.84 terabytes stored in roughly 300 layers within a 12-centimeter-square, 2-millimeter-deep glass chip. That's the equivalent of about 2 million printed books or 5,000 ultrahigh-definition 4K films. (A previous effort in 2019 stored the entirety of the 1978 film *Superman* on a 56-square-centimeter chip.)

Earlier versions of glass data storage often relied on multiple laser pulses to write each voxel. This led to low throughput and high power demands. So scientists developed two different types of voxels, called phase voxels and birefringent voxels, which each required only a single laser pulse to write. By firing millions of laser pulses per second, the system could write up to 25.6 megabits per second with a single beam.

Both phase and birefringent voxels modify the refractive index of the glass, which affects the speed at which light travels through it. A phase voxel alters the refractive index uniformly in all directions, whereas a birefringent voxel changes the refractive index in a direction-dependent manner.

With birefringent voxels, the scientists could achieve significantly better data density, write throughput, and energy efficiency. However, these voxels required high-purity silica glass. In comparison, phase voxels can potentially be written in any durable transparent material, such as borosilicate glass, which is widely available in cookware. Phase voxels also require simpler writing and reading hardware.

To assess the data storage's long-term stability, the scientists repeatedly heated inscribed glass chips up to 500 °C, simulating the long-term aging of the glass at lower temperatures. Their results suggest the data on the plates might be readable for more than 10,000 years at 290 °C, and likely even longer at ambient temperature.

"This is not a replacement for everyday storage like [solid state drives] or hard drives," says Black. "It's designed for data you want to write once and preserve for a very long time." The clearest applications are archival, such as for national libraries, scientific data, or cultural records. ■

AI

A Social Network for AI Agents Reveals Cybersecurity Flaws › The “skills” agents use to function invite vulnerabilities

BY MATTHEW S. SMITH

Moltbook, the first social network for generative AI agents, went live on 28 January, and quickly exploded in popularity. Designed in the style of Reddit, Moltbook is a place where AI agents can post new topics, respond, and upvote or downvote posts autonomously. Agents on Moltbook debate the value of the agent economy, boost cryptocurrencies, and threaten to take over the world, among other topics across more than 1.8 million threads and counting.

Moltbook’s launch led to no shortage of spectacular, and often polarized, headlines. Elon Musk, the CEO of xAI, said it’s the beginning of the singularity; OpenAI CEO Sam Altman called it a fad. There’s one thing that’s not in doubt, however: Agentic AI is a security nightmare.

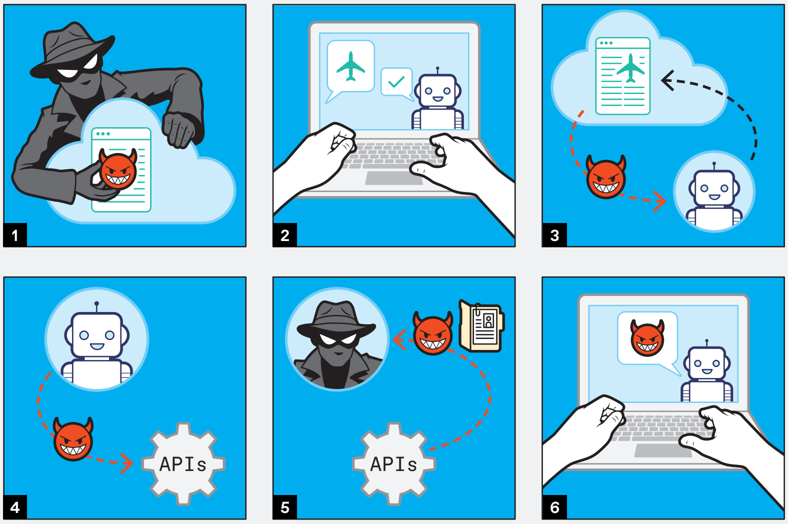
AI security company Snyk found that 36 percent of the codes that provide AI agents their functions contained at least one notable security flaw, and cloud-security company Wiz found a database with open read and write access across all Moltbook data, a flaw that exposed 1.5 million application programming interface (API) keys, which users and programs use to authenticate themselves.

Guillermo Ruiz, a senior solutions architect at AWS (Amazon Web Services), warned that the hype around AI agents can cause people to ignore security issues that might otherwise give them pause. “There’s a lot of people that, with the hype, think ‘I can give my life to it, and just see how it can fix it and solve it,’ ” he says. “But there’s many details behind the scenes that people are not aware of.”

Moltbook is a social network where AI agents can communicate, but it’s not itself an AI agent and has no direct connection to any AI model. It was designed by Matt Schlicht, the CEO of e-commerce company Octane AI, as a place for agents to post and interact. The agents that post on Moltbook are made possible by OpenClaw, an agent framework released by independent software engineer Peter Steinberger. Both OpenClaw and Moltbook were quickly courted by big tech. Steinberger now works for OpenAI, and Moltbook was acquired by Meta for an undisclosed sum.

Whether OpenClaw is itself an AI agent is debatable, as it’s not a chatbot or even an AI model. It’s better understood as server software. It can communicate with hundreds of external services (including Google Search, WhatsApp, and many more) through the WebSocket communications protocol. OpenClaw passes information among these services—which are accessed through code extensions called skills—and an AI model of the user’s choice, like Anthropic’s Claude, Google’s Gemini, or OpenAI’s GPT.

That matters because OpenClaw’s website states it “runs on your machine,” which sounds secure. But while the OpenClaw server can run locally on a wide variety of consumer-grade computer hardware, the majority of people using OpenClaw install skills that communicate with many online services. It’s possible to use OpenClaw agents only inside your own network, and with a locally hosted AI model, but OpenClaw’s own documentation predominately shows it communicating with external services.



How to Trick AI Agents

Malicious actors can indirectly influence the behavior of AI agents by seeding prompts on web pages, in PDFs, or within other materials the agent accesses on the internet.

❶ The attacker hides a malicious prompt in an innocuous online resource. ❷ A user queries an AI agent to complete a task, such as booking plane tickets. ❸ To complete its task, the agent accesses the internet and inadvertently ingests the malicious prompt. ❹ When the agent communicates with other tools (such as APIs) to complete its task, the malicious prompt influences the outcome. ❺ Depending on the tools, the agent could communicate sensitive information back to the attacker. ❻ The agent could also influence the user’s behavior.

Illustration by Chris Philpot

This is why a seemingly benign website like Moltbook can pose a security risk. Snyk's report describes how malicious actors can poison skills that read online data even in situations where the skill itself doesn't contain malicious code. "An attacker can post a prompt-injected message on a forum...and wait for users to invoke the legitimate skill, which faithfully retrieves the poison content." In other words, a third party can change an agent's behavior remotely, and without the user's knowledge, with nothing more than a plain-text prompt on a public website.

Given the risks, the popularity of Moltbook and OpenClaw might seem baffling. Why would anyone choose to use software that can be compromised by accessing a plain-text post online? The answer is simple. OpenClaw can quickly handle tasks that most people find stressful—like buying a car.

AJ Stuyvenberg, a staff engineer at cloud-infrastructure provider Datadog, found himself in the market for a new car just as OpenClaw became popular. Instead of interacting with dealers himself he decided to let an AI agent give it a go. "I asked it to search to find prices and to contact dealerships to see what their best out-the-door price was," says Stuyvenberg. To accomplish that, he gave OpenClaw access to Google Search and email. "That was almost entirely hands-off."

The agent was only able to communicate over email, so when dealers looked to speak over the phone, it gave excuses as to why it wasn't available. And at one point the agent sent an email to the wrong dealer, though this didn't cause any issues in negotiations. After a few days the agent had negotiated a dealer discount of US $4,200.

Although he was happy with the result, Stuyvenberg remains cautious about the security risks this type of agent use creates. "I'm nervous about the scope of what these agents can do, and I've revoked a lot of access, and I give it a much more restricted view over my personal digital life," he says. Still, he was encouraged enough to buy a Mac mini that he intends to use only for OpenClaw. "I thought it was only fair after it saved me all of this money on my car."

The tension between utility and security is a core issue AI agents face, and to see widespread use, they'll need to resolve, or at least mitigate, that tension. As Stuyvenberg's experiment shows, agents can handle tasks that most people would rather not do. Yet they provide the most utility when they have broad access to our digital lives and to online services, which leaves the agent exposed to attack.

AWS's Ruiz argues the issue is broader than any single example of an insecure skill connected to OpenClaw. "I don't think Moltbook is the problem. To be honest, I think the problem is the language—human language," he says. Language is often ambiguous, or open to interpretation. An AI model will likely refuse a direct command to hack a system, but the same model might comply when told that the same command is part of a security audit.

This isn't a flaw specific to OpenClaw's design, Ruiz says, but a consequence of how large language models process instructions. "How do you ensure that nobody is putting a prompt-injection attack inside an email?" He asks. "It's so immense, the kind of things that you need to look after."

While the threats currently outweigh the solutions, OpenClaw is not sitting idle. On 7 February, Steinberger announced that OpenClaw has partnered with cybersecurity service VirusTotal to automatically scan OpenClaw skills. The scans, which are now visible on all skills posted to the official OpenClaw skill library, not only look for malicious code but also for design decisions that can make a skill insecure.

The scans are helpful for catching poorly secured skills, but they're not a full solution. The scans still won't catch prompt-injection attacks because, as mentioned above, these don't exist in the code of the skill itself but instead in the content the skill might access. As there's no obvious way to prevent such attacks, those looking to use AI agents will still need to make careful choices about the services they let agents access. ■

A Prescription for At-Home Brain Stimulation

SOME TECHNOLOGY ENTHUSIASTS **already apply gentle electric current to their brains in an effort to achieve a host of beneficial effects. Now that practice is going more mainstream with Flow Neuroscience's FL-100, which was approved by the U.S. Food and Drug Administration in late 2025 to treat depression. The headset uses transcranial direct current stimulation, which delivers constant low-level DC stimulation via forehead electrodes. Flow already sells the headset without a prescription in the United Kingdom and European Union, and it markets a similar, prescription-free headset for wellness in the United States.** **—Greg Uyeno**

FLOW NEUROSCIENCE

A fleet of Material's Hybrid3D printers could make it possible to manufacture batteries the way office workers print a document.

MATERIAL

TRANSPORTATION

3D Printed Batteries Fit Every Nook and Cranny › Cells to fit any space could extend the range of drones

BY LAWRENCE ULRICH

What do a Formula 1 car, a drone, a soldier's pack, and a wearable smart device have in common? They all need batteries.

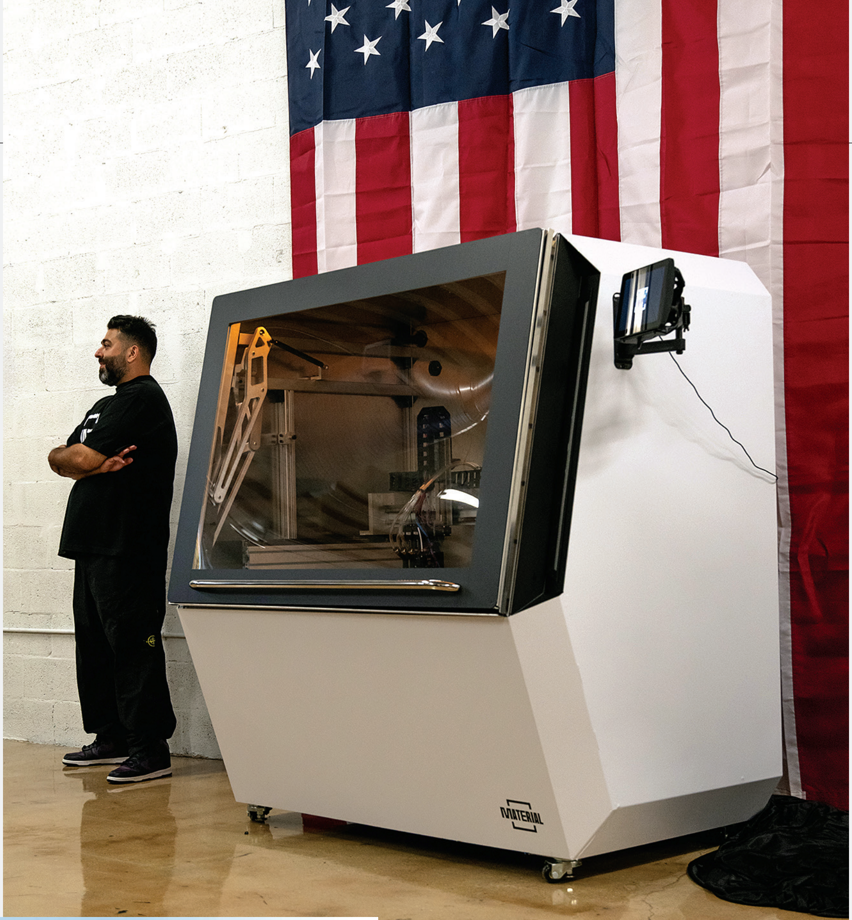
Ideally, those batteries could fit into the oddly shaped spaces within those items, something that today's cylindrical or rectangular cells struggle to do. In 2023, Gabe Elias, who previously helped design Mercedes-AMG Petronas race cars that won seven consecutive F1 constructors' championships, cofounded a startup to 3D print batteries onto surfaces to pack them into those unused spaces in all kinds of devices and vehicles.

The startup, Material Hybrid Manufacturing, recently won a US $1.25 million, 21-month contract with the U.S. Air Force to prove its tech's potential. It joins competitors in a race to popularize printed batteries that can conform to various shapes.

Soon after cofounding the Miami-based company, Elias and his group realized that their initial pitch—printing batteries in new shapes for passenger cars—was stuck in neutral. EVs, especially bigger ones, don't have pressing space constraints for batteries. Electric SUVs and pickup trucks from Rivian, where Elias also worked, can fit 7,776 cylindrical batteries into a 135 kilowatt-hour pack.

So, the company changed lanes to smaller devices. Its Hybrid3D, a proprietary manufacturing platform, can print full-stack batteries in situ: Anode, cathode, separator, and casing, with no molds or costly tooling required. The tech eliminates the metal casings, bus bars, and other components that hog space in conventional cells. The company's active battery material can fill voids and follow three-dimensional curves: Think the wing of a drone, or the slender arm of a pair of smart glasses.

"Things are shrinking, so we're shrinking around [them]," Elias says. "Electronics are becoming embedded,



Material CEO Gabe Elias stands next to one of the company's Hybrid3D printers [above]. Instead of attaching blocky cubes or cylinders, Material can fill the structural elements of a drone with its printed batteries [left].

TOP: MATERIAL; BOTTOM: PERFORMANCE DRONE WORKS

consolidated, optimized, and batteries are the only part of that equation that's being left behind."

The company has teamed up with Performance Drone Works, in Huntsville, Ala., to push its tech toward commercialization. For the initial project with the U.S. Air Force, the companies will show how much active battery their 3D material can pack into the same modular space that holds 48 cylindrical cells in an existing drone. Even in that simplified, proof-of-concept drone, the printed battery is projected to achieve a 50 percent boost in energy density and use 35 percent more available volume.

"That gives you a bunch of options," Elias says. "You could either fly 50 percent farther, or decrease the size of the battery pack, fit more payload, and cover the same distance."

Next-gen designs could boost those gains by dispersing battery material around drone frames, electric motors, or other surfaces. Notoriously heavy military backpacks—now stuffed with bulky, square batteries—could be lighter and ergonomically shaped. Military helmets could directly integrate batteries that power head-mounted gear.

The bed of Material's first commercial-scale printer measures 550 by 350 millimeters, and the company plans to greatly expand that surface. Material's tech is essentially a hybrid of direct ink writing and fused deposition modeling, two of several techniques being developed by companies vying to bring printed energy sources to market.

"Batteries really live and die by layer thicknesses."

—GABE ELIAS, MATERIAL CEO

Critically, the Hybrid3D would allow Material to go from prototyping to printing batteries with no need for expensive, time-consuming retooling. Material's printer platform can already handle a variety of chemistries and formats with simple changes in materials and software coding. "We're chemistry-agnostic," Elias says.

Material's cells currently use liquid polymer electrolyte, added via an infusion process, but the company has a working road map toward solid-state designs. One challenge Material faces is tuning battery materials to flow properly from printer nozzles. That material also must be deposited in uniform, repeatable layers, roughly 50 to 300 micrometers thick, to ensure high quality and yields. "Batteries really live and die by layer thicknesses," Elias says.

He notes that Apple and other companies are investing massive amounts of money to create conformable batteries, such as the L-shaped batteries already used in some iPhones, but are using costly and limited traditional methods. And with consumer electronics giants fighting to popularize wearable devices, printed batteries are an enticing solution for their packaging and power needs.

If companies can manage to print batteries the way office workers print a document, the technology would replace much of the costly tooling, dedicated production lines, and time-consuming processes of conventional batteries. Those printed batteries could then compete on cost across the entire market, Elias says, from single cells to expensive, complex multicell packs.

Material's process has the biggest cost benefit for the most complex battery packs, Elias says. So, when manufacturers face high expenses from those multicell packs, they don't have to cut corners—instead, they can fill them. ■

REACHY MINI

Reachy Mini is an expressive robot designed for creative coding and AI experimentation. Built as a low-cost, accessible platform, it offers a hands-on way to explore robots powered by AI models and supported by a growing open-source robotics community.

CREATOR
Pollen Robotics
(acquired by Hugging
Face in 2025)

YEAR
2025

COUNTRY
France

CATEGORY
Consumer, Research

FEATURES
AI, full autonomy,
modular, open source,
vision

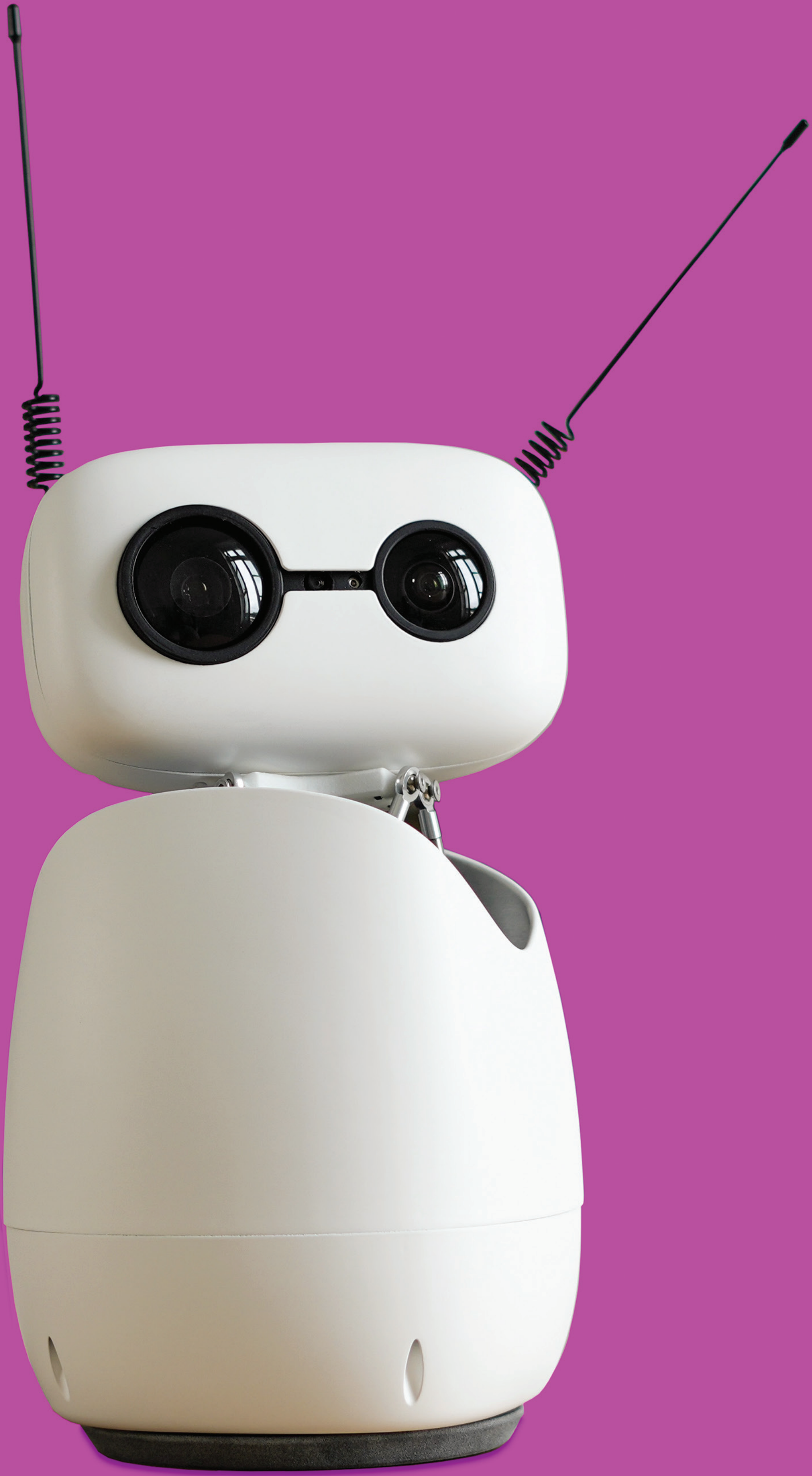
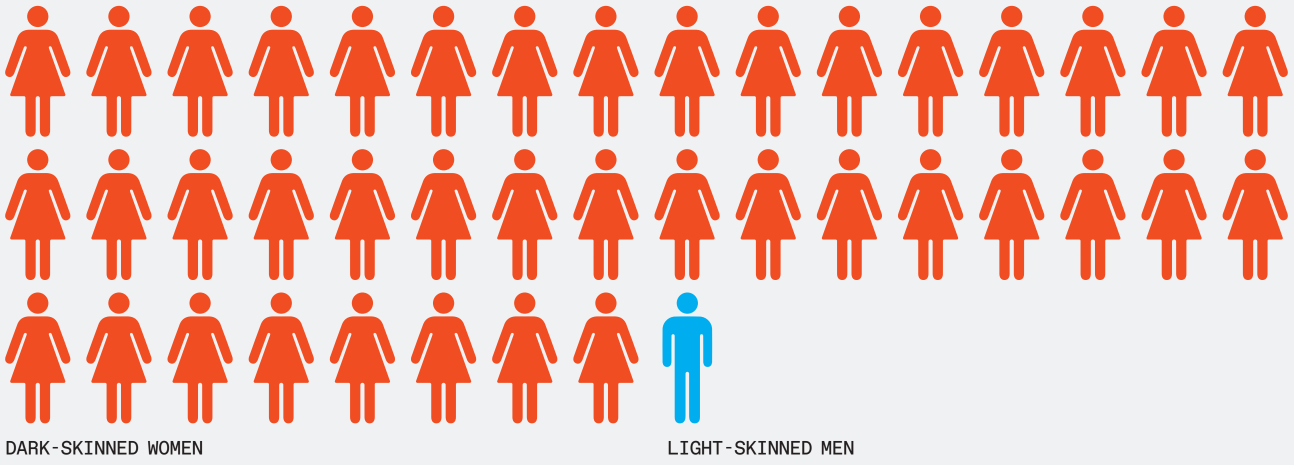


PHOTO: POLLEN ROBOTICS/HUGGING FACE

To learn more about Reachy Mini and explore over 270 real-world robots, with thousands of photos, videos, and interactive animations, scan the QR code or visit the IEEE Robots Guide at robotsguide.com

ROBOTS

The Data



A landmark facial-recognition study in 2018 found that error rates were 40 times as high for dark-skinned women as they were for light-skinned men. Accuracy may have improved since, but real-world uses are creating new problems.

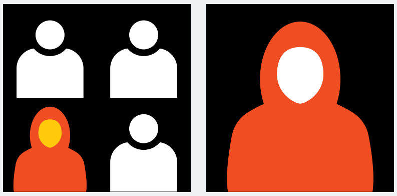
Where Everybody Knows Your Face...

Facial-recognition technology (FRT) dates back 60 years. Just over a decade ago, deep-learning methods tipped the technology into more useful—and menacing—territory. Now, retailers, your neighbors, and law enforcement are all storing your face and building up a fragmentary photo album of your life.

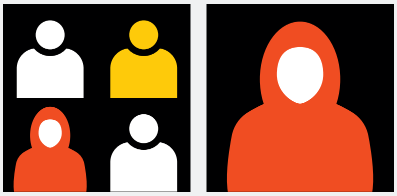
Yet the story those photos can tell inevitably has errors. FRT makers, like those of any diagnostic technology, must balance two types of errors: false positives and false negatives.

In best-case scenarios—such as comparing someone's passport photo to a photo taken by a border agent—false-negative rates are around two in 1,000 and false positives are less than one in 1 million.

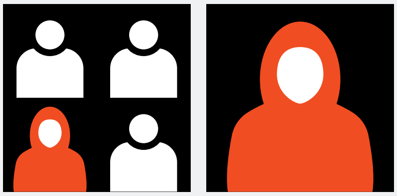
In the rare event you're one of those false negatives, a border agent might ask you to show your passport and take a second look at your face. But as people ask more of the technology, more ambitious applications could lead to more catastrophic errors. Let's say that police are searching for a suspect and they're comparing an image taken with a security camera with a previous "mug shot" of the suspect. The three possible outcomes are shown below.



a) The FRT identifies the suspect, since the two images are of the same person, according to the software. Success!



b) The FRT matches another person in the footage with the suspect's probe image. A false positive, coupled with sloppy verification, could put the wrong person behind bars and lets the real criminal escape justice.



c) The FRT fails to find a match at all. The suspect may be evading cameras, but if the cameras just have low-light or bad-angle images, this creates a false negative. This type of error might let a suspect off and raise the cost of the manhunt.

BY LUCAS LAURSEN

>99%

Accuracy that FRT makers claim for best-case scenarios

EASY TO RECOGNIZE → HARD TO RECOGNIZE

Training-data composition, differences in how sensors detect faces, and intrinsic differences between groups, such as age, all affect an algorithm's performance. The United Kingdom estimated that its FRT exposed some groups, such as women and darker-skinned people, to risks of misidentification as much as two orders of magnitude greater than it did to others.

What happens with photos of people who aren't cooperating, or vendors that train algorithms on biased datasets, or field agents who demand a swift match from a huge dataset? Here, things get murky.

Consider a busy trade fair using FRT to check attendees against a database, or gallery, of images of the 10,000 registrants, for example. Even at 99.9 percent accuracy you'll get about a dozen false positives or negatives, which may be worth the trade-off to the fair organizers. But if police start using something like that across a city of 1 million people, the number of potential victims of mistaken identity rises, as do the stakes.

What if we ask FRT to tell us if the government has ever recorded and stored an image of a given person? That's what U.S. Immigration and Customs Enforcement agents have done since June 2025, using the Mobile Fortify app. The agency conducted more than 100,000 FRT searches in the first six months. The size of the potential gallery is at least 1.2 billion images.

At that size, assuming even best-case images, the system is likely to return around 1 million false matches, but at a rate at least 10 times as high for darker-skinned people, depending on the subgroup.

Responsible use of this powerful technology would involve independent identity checks, multiple sources of data, and a clear understanding of the error thresholds, says computer scientist Erik Learned-Miller of the University of Massachusetts Amherst: "The care we take in deploying such systems should be proportional to the stakes." ■

DETROIT POLICE

THE NEGATIVES OF FALSE POSITIVES

2020: Robert Williams's wrongful arrest cost him detention. The ensuing settlement requires Detroit police to enact policies that recognize FRT's limits.

Security cameras in use

ALGORITHMIC BIAS

2023: Court bans Rite Aid from using facial recognition for five years over its use of a racially biased algorithm.

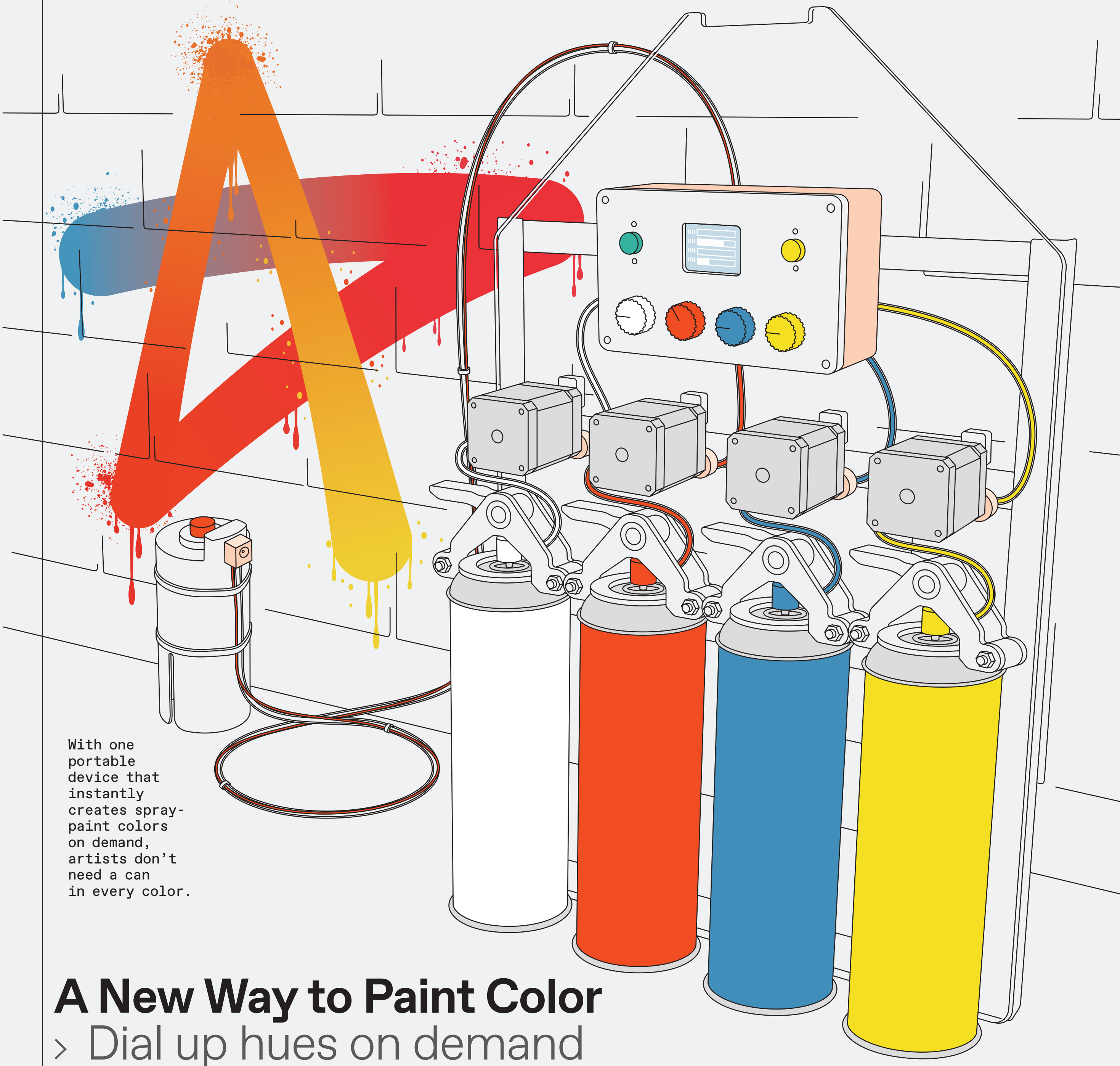
TOO FAST, TOO FURIOUS?

2026: U.S. immigration agents misidentify a woman they'd detained as two different women.

CLOCKWISE FROM BOTTOM RIGHT: VICTOR J. BLUE/BLOOMBERG/GETTY IMAGES; ISTOCK (7)

TECH TO TINKER WITH

Hands On



With one portable device that instantly creates spray-paint colors on demand, artists don't need a can in every color.

A New Way to Paint Color

› Dial up hues on demand

BY SANDESH MANIK

Illustrations by James Provost

We're all familiar with mixing red, yellow, and blue paint in various ratios to instantly make all kinds of colors. This works great for oils or watercolors, but fails when it comes to cans of spray paint. The paint droplets can't be blended once they are aerosolized. Consequently, although spray cans are great for applying even coats of paint to large areas very quickly, spray-paint artists need a separate can for every color they want to use—until now.

Back in 2018, when I first saw professional spray artists lugging dozens to hundreds of cans to their work sites, I was inspired to start noodling on a solution. I've worked at Google X, Alphabet's "moonshot factory," as a hardware engineer, and I'm now building a startup in mechanical-design software. I'm no painter, but I know my way around mechatronics.

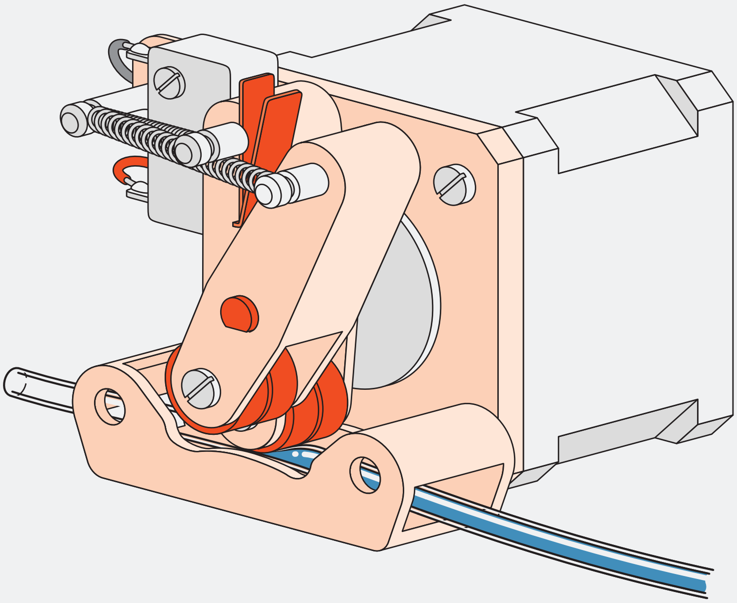
I wanted my solution to be inexpensive and simple enough to build as a DIY project and functional enough for an artist to use without breaking their flow. So I began prototyping a system that combines base colors while they are still in pressurized form from off-the-shelf cans.

I tried a few approaches where pressurized paint from the base-color cans fed through tubes into a mixing channel, before emerging from a spray head. To control the ratios, I decided to borrow a trick that would be familiar to anyone who's ever had to control the brightness of an LED using a microcontroller: pulse-width modulation. Initially, I used electronically controlled solenoid valves to release the paint from the cans. The paint would flow into a mixing channel for relative durations that corresponded to the ratio of the base colors required to make a given hue. However, this failed because different cans never have the same internal pressure. Whenever two valves were open at the same time, the pressure difference would make paint flow backward into the lower-pressure can.

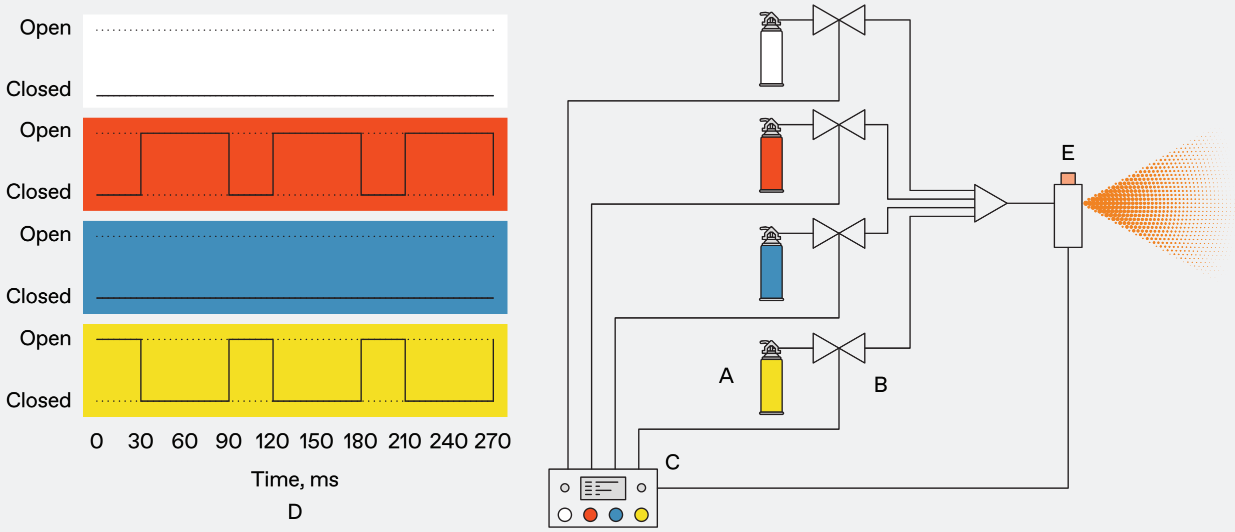
As an alternative, I removed the mixing channel and tried making the paint pulses from each can sequentially converge into a tube so that no more than one valve would ever be open at a time. Surprisingly, this worked perfectly. The backflow was eliminated, and it turned out that the natural turbulence of the flow was sufficient to mix the paints. Let's say you want to produce a clementine orange color. This requires yellow and red paint in a ratio of 1:2, so the yellow valve opens for a period of time, and then the red valve opens for twice as long. The system then keeps repeating this cycle of pulses at a rapid pace to instantly create the spray-paint color you want.

The theory is straightforward, but making this work in practice took quite a bit of experimentation. First, I had to determine the actual durations of pulses that would produce evenly mixed colors, not just their ratios. I also needed to work out the size of the tubing (too narrow and you'd get low spray force; too wide and you'd have paint accumulating in the tubes). Eventually I settled on a maximum pulse duration of 250 milliseconds and a tube diameter of 1 millimeter.

Even though the system worked, the solenoid valves I used constantly clogged up. Designed for water purifiers, the valves didn't prevent paint from entering the mechanism, where the paint would harden. Moreover, when the valves



This new rotary pinch-valve design can open and close rapidly, eliminating the problem of back pressure when mixing paints.



Cans holding base colors (A) are attached to valves (B). An Arduino-based control panel (C) opens and closes valves to mix paint before it is aerosolized (E). By quickly opening and closing valves with varying durations in sequence (D), you can mix paint in specific ratios to create desired colors.

were turned off, they could stop backflow only if the inlet remained pressurized. So disconnecting a paint can from the system would cause instant leaking. Other off-the-shelf valves I tried couldn't cycle fast enough and were too expensive.

So I created my own mechanism: a high-speed, electronically controlled, rotary pinch valve. It has a stepper motor that rotates a lever with a rolling bearing to constrict fluid flow inside a flexible tube. This concept isn't new—there's something like them in every peristaltic pump. But I added a spring to firmly hold the lever in the closed position against any back pressure when the motor isn't powered, making it a normally closed valve that isolates the attached can. Additionally, the valve is fast enough to be open for as little as 30 ms.

I went through four major prototypes of the system before reaching a working version, and I had some spectacular failures along the way of the sort that only pressurized paint can provide. The final version uses four base colors—red, yellow, blue, and white—with the color mix controlled by four knobs attached to an Arduino Nano and a small display. The flow of paint is triggered by a push button placed above the spray head, similar to a spray can's nozzle.

> **I had some spectacular failures along the way of the sort that only pressurized paint can provide.**

The length of time a base color's paint valve can be open is one of eight values between 30 and 250 ms. This means that the entire system—which I coincidentally dubbed Spectrum—can create hundreds of distinct spray-paint colors instantly. It produces fewer than 8^4 (or 4,096) colors because duration ratios that are a multiple of each other will produce the same color—for example, 2:3 and 4:6. I added a force sensor to the push button, which allows for a gradient: Two color mixes can be dialed in, and as I increase my thumb's pressure on the button, the paint mix shifts from one color to the other.

Spectrum's various fixtures are 3D-printed, and project details and videos are available through my website at https://www.sandeshmanik.com/projects/spectrum. Preprints of technical descriptions of the rotary pinch valve and mixing methodology are available on TechRxiv. The total cost for the bill of materials is less than US $150.

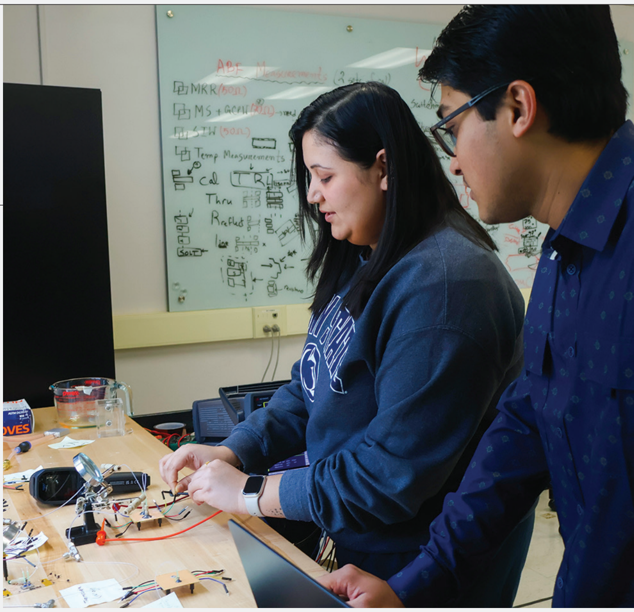
Working on and off on the side for about seven years, I finally finished developing my system and writing the documentation in late 2025. After I posted a video to social media, I was heartened by the immediate positive response from spray-paint artists around the world. I'm now creating step-by-step instructions so that nontechnical people can build their own Spectrum paint sprayer. I look forward to seeing what creations artists out in the wild make! ■

BY AARON MOK

Careers

U.S. Engineering Ph.D. Programs Are Under Strain

Enrollment is down amid funding and immigration uncertainty



Two Ph.D. students at Penn State, Ruposri Bhattacharjee and Md Sultan Mahmud, discuss a microfluidics test setup. Both students are advised by CHIMES director Madhavan Swaminathan.

MAY CREWS/PENN STATE

U.S. doctoral programs in electrical engineering form the foundation of technological advancement, training the brightest minds in the world to research, develop, and design next-generation electronics, software, electrical infrastructure, and other high-tech products and systems. Elite institutions have long served as launchpads for the engineers behind tomorrow's technology.

Now that foundation is under strain.

With U.S. universities increasingly entangled in political battles under the second Trump administration, uncertainty is beginning to ripple through doctoral admissions for electrical engineering programs. While some departments are reducing the number of spots available in anticipation of potential federal funding cuts, others are seeing their applicant pools shrink, particularly among international students, who make up a significant portion of their programs.

The number of computing Ph.D.s grew significantly in the 2010s, according to data from the National Academies, but there is still high demand for those with advanced degrees across academia, government, and industry. Now, some universities point to warning signs of waning enrollment.

Though not all engineers have Ph.D.s, if enrollment continues to shrink, fewer doctoral students could mean fewer engineers developing cutting-edge technology and training the next generation, potentially exacerbating existing labor shortages as global competition for tech talent intensifies.

Public universities in particular are feeling the strain because they rely heavily on federal grants to support doctoral students.

The University of California, Los Angeles, for instance, must fund Ph.D. students for the duration of a degree—typically five years. In August 2025, the U.S. government pulled more than US $580 million in federal grants over allegations that the university failed to adequately address antisemitism on campus during student protests. A federal judge has since ordered the funding to be restored, but faculty began to worry that research support could be clawed back without notice, says Subramanian Iyer, distinguished professor at UC Los Angeles's department of electrical and computer engineering.

According to Iyer, departments across UC Los Angeles, including engineering, plan to scale back Ph.D. admissions this year. "The fear is that at some point, all this government money will be taken away," Iyer says. "Lowering the admissions rate is just a way to prepare for that reality."

In response to a request for comment, a spokesperson for the U.S. National Science Foundation—a major source of federal research funding at UC Los Angeles and elsewhere—said, "NSF recognizes the essential role doctoral trainees play in the nation's engineering and STEM enterprise."

Funding shocks may also force Pennsylvania State University to reshape future admissions decisions, according to Madhavan Swaminathan, head of Penn State's electrical engineering department and director of the Center for Heterogeneous Integration of Micro Electronic Systems (CHIMES), a semiconductor research lab.

In 2023, the Defense Advanced Research Projects Agency (DARPA) and industry partners awarded CHIMES a five-year $32.7 million grant. But in late 2025, the agency pulled its final year of funding from the center, citing a shift in priorities from microelectronics to photonics, Swaminathan

> **“The fear is that at some point, all this government money will be taken away.”**
>
> —Subramanian Iyer, distinguished professor, UC Los Angeles

says. As a result, CHIMES’s annual budget, which supports research assistantships for roughly 100 engineering graduate students, the majority pursuing Ph.D.s, will fall from $7 million in 2026 to $3.5 million in 2027. If these constraints persist, Penn State’s engineering department may reduce the number of doctoral students it supports.

In a statement, a DARPA spokesperson told *IEEE Spectrum,* “Basic research is central to identifying world-changing technologies, and DARPA remains committed to engaging academic institutions in our program research. By design, a DARPA program typically lasts about three to five years. Once we establish proof of concept, we transition the technology for further development and turn our attention to other challenging areas of research.”

Penn State’s enrollment numbers reflect Swaminathan’s caution. He says the electrical engineering Ph.D. cohort shrank from 28 students in 2024 to 15 students in 2025. Applications show a similar pattern—a sign that prospective students may be wary of applying to U.S. programs.

In late January, the Trump administration announced it had paused visa approvals for the citizens of 75 countries. Months earlier, the administration proposed new restrictions on student visas, including a four-year cap.

For Texas A&M University’s graduate electrical and computer engineering programs, up to 80 percent of applicants each year are international students, according to Narasimha Annapareddy, professor and head of the department. Annapareddy says applications for the fall 2026 Ph.D. cohort have dropped by roughly 50 percent.

Annapareddy says the United States is “sending a message that migration is going to be more difficult in the future.” Foreign students often pursue degrees in the U.S. not only for academic training, he says, but to build long-term careers and lives in the country. Fewer applications from international students mean that the university forgoes a “driven and hungry” segment of the applicant pool who are highly qualified in technical fields.

At the University of Southern California, the decline is more moderate. The freshman Ph.D. class fell from about 90 students in 2024 to roughly 70 in 2025, a reduction of 22 percent, according to Richard Leahy, department chair of USC’s Ming Hsieh Department of Electrical and Computer Engineering.

While Leahy says applications are down modestly overall, domestic applications have increased by roughly 15 percent. Beyond immigration restrictions, international students, particularly from countries such as India and China, may be staying in their home countries as their technology sectors expand.

“A lot of those students that would normally have come to the U.S. are now taking very good jobs working in the AI industry and other areas,” Leahy says. “There are a lot more opportunities now.”

Some faculty say shrinking cohorts could erode the tech workforce if the pattern continues.

At UC Los Angeles, Iyer describes a doctoral ecosystem built on a chain of mentorship. Among the roughly 25 students in his lab, senior doctoral students mentor junior Ph.D. candidates, who in turn guide master’s students and undergraduates. The system depends on overlapping cohorts. Reducing the number of students hired weakens that overlap and the trickle-down benefits of the mentorship model that keeps labs functioning.

The real benefit of the university system isn’t just the teaching but also “the community that you build,” Iyer says. “As you decrease admissions, this will disappear.”

At Penn State, Swaminathan points to specialization as key to a strong workforce. Many doctoral students train in semiconductor engineering, feeding expert talent into the domestic chip industry. If enrollment continues to shrink over the next few years, Swaminathan says, companies may need to hire students with bachelor’s or master’s degrees, who might lack the necessary skills required to design and innovate new chips.

“Without that specialization, there’s only so much one can do,” he says.

Not all departments are shrinking. At the University of Texas at Austin, overall enrollment has remained relatively steady, according to Diana Marculescu, chair of UT Austin’s Chandra Family Department of Electrical and Computer Engineering.

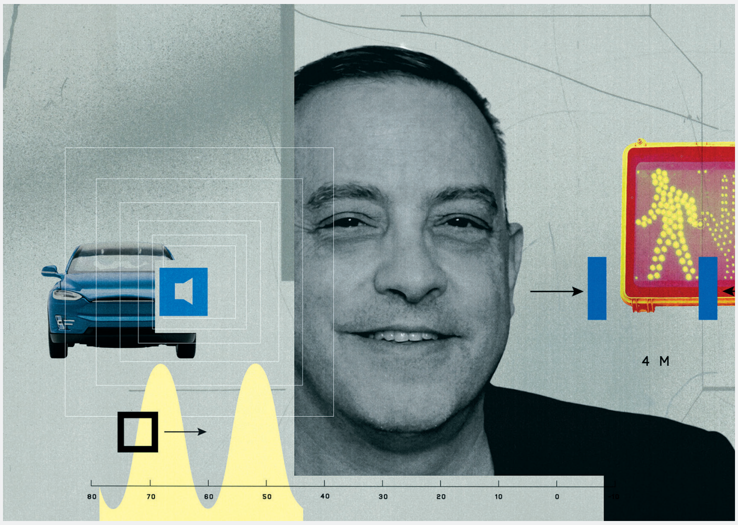
While she says recent fluctuations aren’t raising alarms, her concern lies more with alignment between research and industry. Doctoral students often train according to current grant priorities, she says. But by the time graduates enter the job market four to six years later, their specialization may not align neatly with open roles. That creates friction in the talent pipeline.

“That lack of connection might be problematic,” Marculescu says. She argues that closer collaboration between universities and the private sector could help create stronger feedback loops between hiring needs and academic research priorities.

For now, USC’s Leahy says Ph.D. graduates remain in high demand, and the current shifts have not yet translated into measurable workforce shortages. “We should be concerned about the number of Ph.D.s,” he says. “But there isn’t a crisis at this point.” ■

BY AARON MOK

5 Questions



Michael Roan

EV sounds are a surprisingly complicated science

As more EVs hit the roads, pedestrians face a challenge: hearing them. Regulators in the United States, Europe, and elsewhere require EVs to emit warning sounds at low speeds. What may seem like a simple safety measure is a complex web of acoustics, signal design, and human perception.

BMW, Hyundai, Mercedes-Benz and others have enlisted sound designers, acoustics specialists, and even musicians to craft sounds for their cars. *IEEE Spectrum* spoke with Michael Roan, who has researched how artificial vehicle sounds are perceived, to learn more about the challenges.

What do automakers optimize EV sounds for?

Michael Roan: The No. 1 priority is safety. After that, it's brand identity. Where they clash is in trying to make a sound that helps the brand stand out but still satisfies government regulations.

Regulations specify the frequency content of the signal, meaning certain frequency bands have to contain specific levels of energy. That ends up being a very tonal signal that isn't super aesthetically pleasing. So car manufacturers have to add sound at different frequency bands to develop something that's identifiable as their brand, while still satisfying those regulations.

Michael Roan is a professor of mechanical engineering at Penn State University. His research includes how artificial vehicle sounds are perceived, particularly by visually impaired pedestrians.

How do automakers design the audio systems?

Roan: Getting the sound out to the pedestrian with just one little speaker is very difficult because it's not a direct acoustic path. The sound bounces around inside the vehicle and off the road. That creates constructive and destructive interference, meaning there are certain spots where you don't hear the signal at all.

EVs are designed with a single speaker because it's the cheapest, simplest way to meet safety regulations. Ensuring that the speaker produces uniform loudness around the car is a challenge. From a hardware perspective, it may actually be advantageous to have a more distributed sound source, like multiple speakers.

How well do warning sounds work in real-world conditions?

Roan: Nobody knows. There hasn't been a real set of tests done with human listeners who are sitting there doing perception tests in an urban, complicated environment. They're done in a very quiet area with background noise that's pumped in via loudspeakers under very controlled conditions.

Do EV sound regulations improve safety?

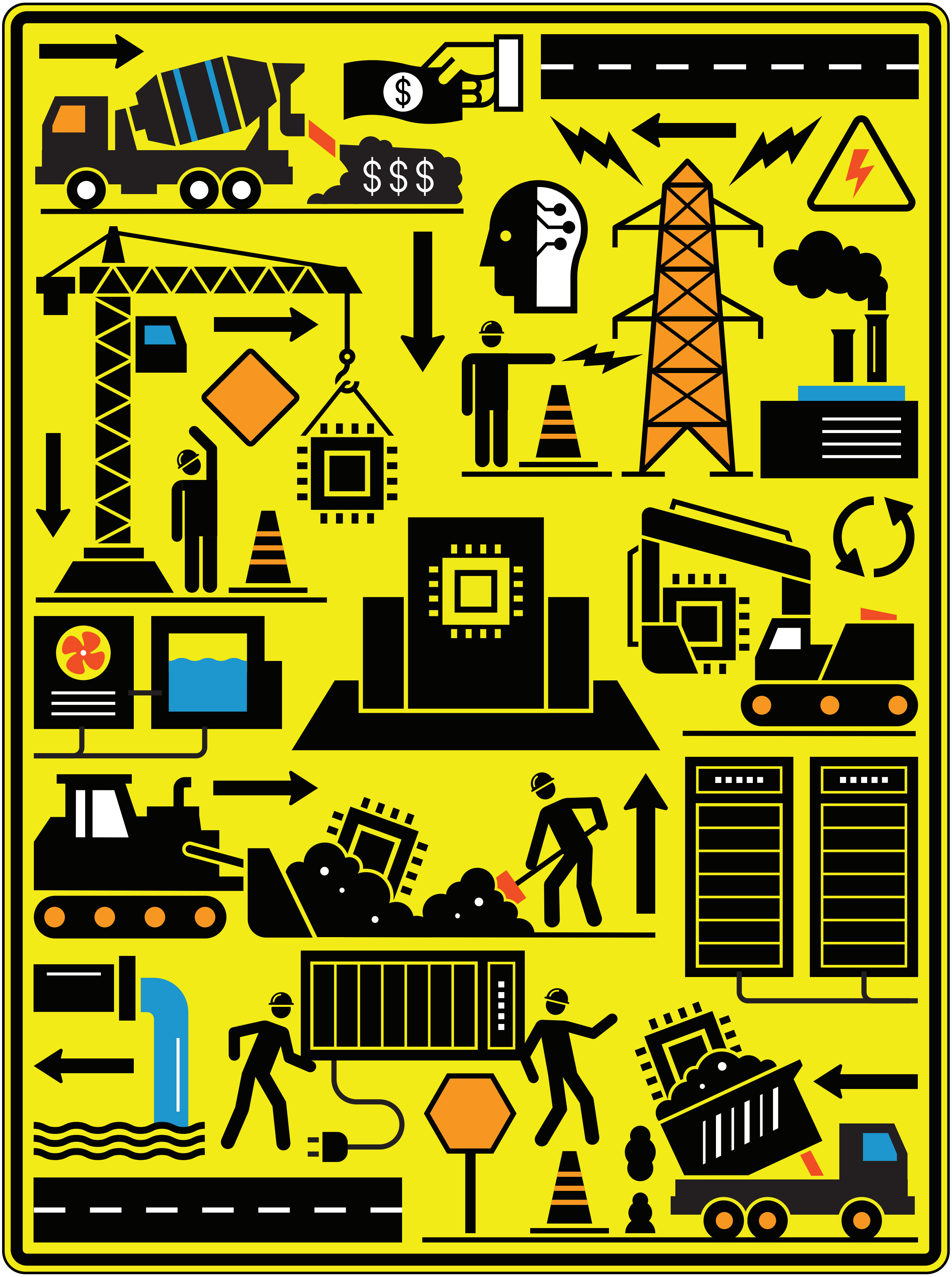
Roan: They absolutely improve safety. Testing done by us and others shows that if you compare an electric vehicle with no added sound to one with a sound that satisfies the requirements, there is a huge jump in the probability of detection by a large group of visually impaired pedestrians.

Should EVs sound like cars with internal combustion engines, or something new?

Roan: An internal combustion engine is large and has many moving parts. There's an exhaust at the back of the vehicle that creates a huge distributed sound. And you're saying, "Oh, I'm going to pump that through a little bitty 4-inch speaker that comes out through the wheel well." It's just not perceptible.

I think the manufacturers have to start with meeting the regulations. Then they can go beyond that to sounds that are "vehicle-like," although that's very subjective. Once the regulations are met, then artistry can come into play to develop brand identity. ■

Photo-illustration by Stuart Bradford



WHAT WILL IT TAKE TO BUILD THE WORLD'S LARGEST DATA CENTER?

META'S 5-GIGAWATT AI DATA CENTER IS FORCING ENGINEERS TO THROW OUT THE RULE BOOK

By Matthew S. Smith
Illustrations by Optics Lab

The undying thirst for smarter (historically, that means larger) AI models and greater adoption of the ones we already have has led to an explosion in data-center construction projects, unparalleled both in number and scale. Chief among them is Meta's planned 5-gigawatt data center in Louisiana, called Hyperion, announced in June of 2025. Meta CEO Mark Zuckerberg said Hyperion will "cover a significant part of the footprint of Manhattan," and the first phase—a 2-GW version—will be completed by 2030.

Though the project's stated 5-GW scale is the largest among its peers, it's just one of several dozen similar projects now underway. According to Michael Guckes, chief economist at construction-software company ConstructConnect, spending on data centers topped US $27 billion by July of 2025 and, once the full-year figures are tallied, will easily exceed $40 billion. Hyperion alone accounts for about a quarter of that.

For the engineers assigned to bring these projects to life, the mix of challenges involved represent a unique moment. The world's largest tech companies are opening their wallets to pay for new innovations in compute, cooling, and network technology designed to operate at a scale that would've seemed absurd five years ago.

At the same time, the breakneck pace of building comes paired with serious problems. Modern data-center construction frequently requires an influx of temporary workers and sharply increases noise, traffic, pollution, and often local electricity prices. And the environmental toll remains a concern long after facilities are built, due to the unprecedented 24/7 energy demands of AI data centers which, according to one recent study, could emit the equivalent of tens of millions of tonnes of CO_2 annually in the United States alone.

Regardless of these issues, large AI companies, and the engineers they hire, are going full steam ahead on giant data-center construction. So, what does it really take to build an unprecedentedly large data center?

The stereotypical data-center building rests on a reinforced concrete slab foundation. That's paired with a steel skeleton and poured concrete wall panels. The finished building is called a "shell," a term that implies the structure itself is a secondary concern. Meta has even used gigantic tents to throw up temporary data centers.

Still, the scale of the largest AI data centers brings unique challenges. "The biggest challenge is often what's under the surface. Unstable, corrosive, or expansive soils can lead to delays and require serious intervention," says Robert Haley, vice president at construction consulting firm Jacobs. Amanda Carter, a senior technical lead at Stantec, said a soil's thermal conductivity is also important, as most electrical infrastructure is placed underground. "If the soil has high thermal resistivity, it's going to be difficult to dissipate [heat]." Engineers may take hundreds or thousands of soil samples before construction can begin.

There's apparently no shortage of eligible sites, however, as both the number of data centers under construction, and the money spent on them, has skyrocketed. The spending has allowed companies building data centers to throw out

GPUs

MODERN AI DATA CENTERS often use rack-scale systems, such as the Nvidia GB200 NVL72, which occupy a single data-center rack. Each rack contains 72 GPUs, 36 CPUs, and up to 30.4 terabytes of memory. The racks measure over 2.2 meters tall and weigh over one and a half tonnes, forcing AI data centers to use thicker concrete with more reinforcement to bear the load.

A SINGLE GB200 RACK can use up to 120 kilowatts. If Hyperion meets its 5-gigawatt goals, the data-center campus could include over 41,000 rack-scale systems, for a total of more than 3 million GPUs. The final number of GPUs used by Hyperion is likely to be less than that, though only because future GPUs will be larger, more capable, and use more power.

MONEY

▲ **ACCORDING TO CONSTRUCTCONNECT,** spending on data centers neared US $27 billion through July of 2025 and, according to the latest data, will tally close to $60 billion through the end of the year. Meta's Hyperion project is a big slice of the pie, at $10 billion.

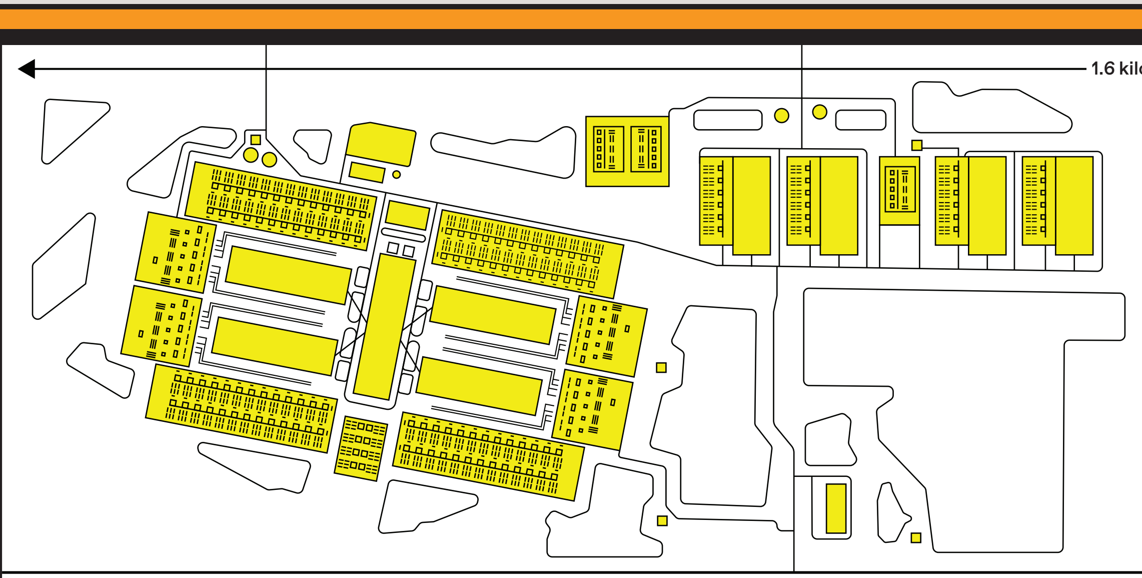
▲ ▲ **DATA CENTER SPENDING** has become an important prop for the construction industry, which is seeing reduced demand in other areas, such as residential construction and public infrastructure. ConstructConnect's third quarter 2025 financial report stated that the quarter's decline "would have been far more severe without an $11 billion surge in data center starts."

MEMORY

IN ADDITION TO RAW COMPUTE PERFORMANCE, Nvidia GB200 NVL72 racks also require huge amounts of memory. An Nvidia GB200 NVL72 rack may include up to 13.4 terabytes of high-bandwidth memory, which implies a data-center campus at Hyperion's scale will require at least several dozen petabytes.

THE IMMENSE DEMAND has sent memory prices soaring: The price of DRAM, specifically DDR5, has increased 172 percent in 2025.

LAND



→ **META CEO MARK ZUCKERBERG** kicked off the buzz around Hyperion by saying it would cover a large chunk of Manhattan. Many took that to mean Hyperion would be a single building of that size, which isn't correct. Hyperion will actually be a cluster of data centers—11 are currently planned—with over 370,000 square meters of floor space. That's a lot smaller even than New York City's Central Park, which covers 6 percent of Manhattan. Meta has room to grow, however. The Richland Parish site spans 14.7 million square meters in total, which is about a quarter the area of Manhattan. And the 370,000 square meters of floor space Hyperion is expected to provide doesn't include external infrastructure, such as the three new combined-cycle gas power plants Louisiana utility Entergy is building to power the project.

POWER

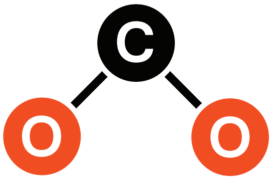
① **HYPERION IS EXPECTED TO** use 5 gigawatts of power across 11 buildings, which works out to just under 500 megawatts per building, assuming each will be similar to its siblings. That's enough to power roughly 4.2 million U.S. homes.

② **JUST ONE HYPERION** data center built at the Richland Parish site will consume twice as much power as xAI's Colossus which, at the time of its completion in the summer of 2024, was among the largest data centers yet built.

CO$_2$

▲ **RESEARCH PUBLISHED IN *NATURE*** in 2025 projects that data-center emissions will range from 24 million to 44 million CO_2-equivalent tonnes annually through 2030 in the United States alone. While some materials used in data centers, such as concrete, lead to significant emissions, the majority of these emissions will result from the high energy demands of AI servers.

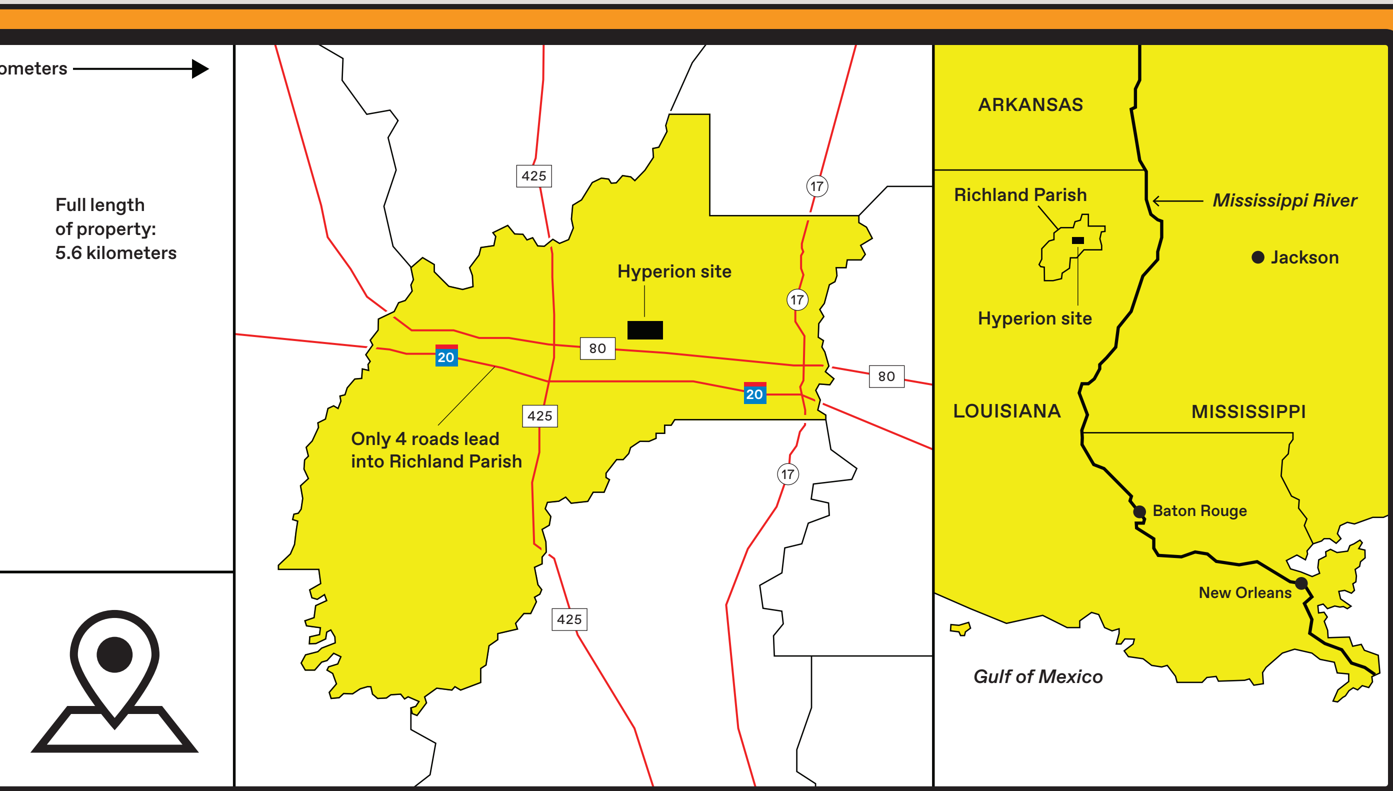
▲ **ESTIMATING THE CARBON EMISSIONS** of Hyperion is difficult, as the project won't be completed until 2030. Assuming that the three new natural gas plants that are planned for construction as part of the project produce emissions typical for their type, however, the plants could lead to full life-cycle emissions of between 4 million and 10 million tonnes of CO_2 annually—roughly equivalent to the emissions of a country like Latvia.



CONCRETE

● **DATA CENTERS ARE TYPICALLY** built from concrete, with steel used as a skeleton to reinforce and shape the concrete shell. While the foundation is often poured concrete, the walls and floors are most often built from prefabricated concrete panels that can span up to 23 meters. Floors use a reinforced T-shape, similar to a steel girder, measuring up to 1.2 meters across at its thickest point. The largest data centers include hundreds of these concrete panels.

● **THE AMERICA CEMENT ASSOCIATION** projects that the current surge in building will require 1 million tonnes of cement over the next three years, though that's still a tiny fraction of the overall cement industry, which weighed in at roughly 103 million tonnes in 2024.


ometers
Full length of property: 5.6 kilometers
425
17
Hyperion site
17
80
20
80
20
425
Only 4 roads lead into Richland Parish
17
425
ARKANSAS
Richland Parish
Mississippi River
Jackson
Hyperion site
LOUISIANA
MISSISSIPPI
Baton Rouge
New Orleans
Gulf of Mexico

the rule book. Prior to the AI boom, most data centers relied on tried-and-true designs that prioritized inexpensive and efficient construction. Big tech's willingness to spend has shifted the focus to speed and scale.

The loose purse strings open the door to larger and more robust prefabricated concrete wall and floor panels. Doug Bevier, director of development at Clark Pacific, says some concrete floor panels may now span up to 23 meters and need to handle floor loads up to 3,000 kilograms per square meter, which is more than twice the load international building codes normally define for manufacturing and industry. In some cases, the concrete panels must be custom-made for a project, an expensive step that the economics of pre-AI data centers rarely justified.

Simultaneously, the time scale for projects is also compressed: Jamie McGrath, senior vice president of data-center operations at Crusoe, says the company is delivering projects in "about 12 months," compared to 30 to 36 months before. Not all projects are proceeding at that pace, but speed is universally a priority.

That makes it difficult to coordinate the labor and materials required. Meta's Hyperion site, located in rural Richland Parish, La., is emblematic of this challenge. As reported by NOLA.com, at least 5,000 temporary workers have flocked to the area, which has only about 20,000 permanent residents. These workers earn above-average wages and bring a short-term boost for some local businesses, such as restaurants and convenience stores. However, the project has also spurred complaints from residents about traffic, construction noise, and pollution.

This friction with residents includes not only these obvious impacts, but also things you might not immediately suspect, such as light pollution caused by around-the-clock schedules. Also significant are changes to local water tables and runoff, which can reduce water quality for neighbors who rely on well water. These issues have motivated a few U.S. cities to enact data-center bans.

Meta's Richland Parish site also highlights a problem that's priority No. 1 for both AI data centers and their critics: power.

Data centers have always drawn large amounts of power, which nudged data-center construction to cluster in hubs where local utilities were responsive to their demands. Virginia's electric utility, Dominion Energy, met demand with agreements to build new infrastructure, often with a focus on renewable energy.

The power demands of the largest AI data centers, though, have caught even the most responsive utilities off guard. A report from the Lawrence Berkeley National Laboratory, in California, estimated the entire U.S. data-center industry consumed an average load of roughly 8 GW of power in 2014. Today, the largest AI data-center campuses are built to handle up to a gigawatt each, and Meta's Hyperion is projected to require 5 GW.

"Data centers are exasperating issues for a lot of utilities," says Abbe Ramanan, project director at the Clean Energy Group, a Vermont-based nonprofit.

Ramanan explains that utilities often use "peaker plants" to cope with extra demand. They're usually older, less efficient fossil-fuel plants which, because of their high cost to operate and carbon output, were due for retirement. But Ramanan says increased electricity demand has kept them in service.

Meta secured power for Hyperion by negotiating with Entergy, Louisiana's largest electricity generator, for construction of three new gas-turbine power plants. Two will be located near the Richland Parish site, while a third will be located in southeast Louisiana.

Entergy frames the new plants as a win for the state. "A core pillar of Entergy and Meta's agreement is that Meta pays for the full cost of the utility infrastructure," says Daniel Kline, director of power-delivery planning and policy at Entergy. The utility expects that "customer bills will be lower than they otherwise would have been." That would prove an exception, as a recent report from Bloomberg found electricity rates in regions with data centers are more likely to increase than in regions without.

The plants, which will generate a combined 2.26 GW, will use combined-cycle gas turbines that recapture waste heat from exhaust. This boosts thermal efficiency to 60 percent and beyond, meaning more fuel is converted to useful energy. Simple-cycle turbines, by contrast, vent the exhaust, which lowers efficiency to around 40 percent.

Even so, total life-cycle emissions for the Hyperion plants could range from 4 million to over 10 million tonnes of CO_2 each year, depending on how frequently the plants are put in use and the final efficiency benchmarks once built. On the high end, that's as much CO_2 as produced by over 2 million passenger cars. Fortunately, not all of Meta's data centers take the same approach to power. The company has

Meta's Hyperion data center is under construction in Richland Parish, La., on a sprawling site about a quarter the area of Manhattan.

META

announced a plan to power Prometheus, a large data-center project in Ohio scheduled to come online before the end of 2026, with nuclear energy.

Other big tech companies, spurred by the need to build data centers quickly, are taking a less efficient approach. xAI's Colossus 2, located in Memphis, is the most extreme example. The company trucked dozens of temporary gas-turbine generators to power the site located in a suburban neighborhood. OpenAI, Oracle, and Softbank, meanwhile, have gas turbines capable of generating up to 300 megawatts at their new Stargate data center in Abilene, Texas, slated to open later in 2026. Both use simple-cycle turbines with a much lower efficiency rating than the combined-cycle plants Entergy will build to power Hyperion.

Demand for gas turbines is so intense, in fact, that wait times for new turbines are up to seven years. Some data centers are turning toward refurbished jet engines to obtain the turbines they need.

The demand for new, reliable power is driven by the power-hungry GPUs inside modern AI data centers.

In January of 2025, Mark Zuckerberg announced in a post on Facebook that Meta planned to end 2025 with at least 1.3 million GPUs in service. Stargate, OpenAI's joint venture, plans to use over 450,000 Nvidia GB200 GPUs, and xAI's Colossus 2, an expansion of Colossus, is built to accommodate over 550,000 GPUs.

GPUs, which remain by far the most popular for AI workloads, are bundled into human-scale monoliths of steel and silicon which, much like the data centers built to house them, are rapidly growing in weight, complexity, and power consumption.

Nvidia's GB200 NVL72—a rack-scale system—is currently a leading choice for AI data centers. A single GB200 rack contains 72 GPUs, 36 CPUs, and up to 30.4 terabytes of memory. It measures 2.2 meters tall, tips the scales at up to 1,553 kilograms, and consumes about 120 kilowatts—as much as around 100 U.S. homes. And this, according to Nvidia, is just the beginning. The company anticipates future racks could consume up to a megawatt each.

Viktor Petik, senior vice president of infrastructure solutions at Vertiv, says the rapid change in rack-scale AI systems has forced data centers to adapt. "AI racks consume far more power and weigh more than their predecessors," says Petik. He adds that data centers must supply racks with multiple power feeds, without taking up extra space.

The new power demands from rack-scale systems have consequences that are reflected in the design of the data center—even its footprint.

In 2022 Meta broke ground on a new data center at a campus in Temple, Texas. According to SemiAnalysis, which studies AI data centers, construction began with the intent to build the data center in an H-shaped configuration common to other Meta data centers.

Construction was paused midway in December of 2022, however, as part of a company-wide review of its data-center infrastructure. Meta decided to knock down the structure it had built and start from scratch. The reasons for this decision were never made public, but analysts believe it was due to the old design's inability to deliver sufficient electricity to new, power-hungry AI racks. Construction resumed in 2023.

Meta's replacement ditches the H-shaped building for simple, long, rectangular structures, each flanked by rows of gas-turbine generators. While Meta's plans are subject to change, Hyperion is currently expected to comprise 11 rectangular data centers, each packed with hundreds of thousands of GPUs, spread across the 13.6-square-kilometer Richland Parish campus.

Nvidia's ultradense AI GPU racks are changing data centers not only with their weight, and power draw, but also with their intense cooling and bandwidth requirements.

Data centers traditionally use air cooling, but that approach has reached its limits. "Air as a cooling medium is inherently inferior," says Poh Seng Lee, head of CoolestLAB, a cooling research group at the National University of Singapore.

Instead, going forward, GPUs will rely on liquid cooling. However, that adds a new layer of complexity. "It's all the way to the facilities level," says Lee. "You need pumps, which we call a coolant distribution unit. The CDU will be connected to racks using an elaborate piping network." On the rack, pipes connect to cold plates mounted atop every GPU; outside the data-center shell, pipes route through evaporation cooling units. Lee says retrofitting an air-cooled data center is possible but expensive.

The networking used by AI data centers is also changing to cope with new requirements. Traditional data centers were positioned near network hubs for easy access to the global internet. AI data centers, though, are more concerned with networks of GPUs. These connections must sustain high bandwidth with impeccable reliability.

This is a point where the scale of Meta's Hyperion, and other large AI data centers, can be deceptive. It seems to imply the physical size of a single data center is what matters. But rather than being a single building, Hyperion is actually a set of buildings connected by high-speed fiber optics.

"Interconnecting data centers is absolutely essential," says Mark Bieberich, a vice president at network-infrastructure company Ciena. "You could think about it as one logical AI training facility, but with geographically distributed facilities." Nvidia has taken to calling this "scale across," to contrast it with the idea that data centers must "scale up" to larger singular buildings.

The full scale of the challenges that face Hyperion, and other future AI data centers of similar scale, remain hazy. Nvidia has yet to introduce the rack-scale AI GPU systems it will host. How much power will it demand? What type of cooling will it require? How much bandwidth must be provided? These can only be estimated.

In the absence of details, the gravity of AI data-center design is pulled toward one certainty: It must be big. New data-center designers are rewriting their rule book to handle power, cooling, and network infrastructure at a scale that would've seemed ridiculous five years ago.

This innovation is fueled by big tech's fat wallet, which shelled out tens of billions of dollars in 2025 alone, leading to questions about whether the spending is sustainable. For the engineers in the trenches of data-center design, though, it's viewed as an opportunity to make the impossible possible.

"I tell my engineers, this is peak. We're being engineers. We're being asked complicated questions," says Stantec's Carter. "We haven't got to do that in a long time." ■

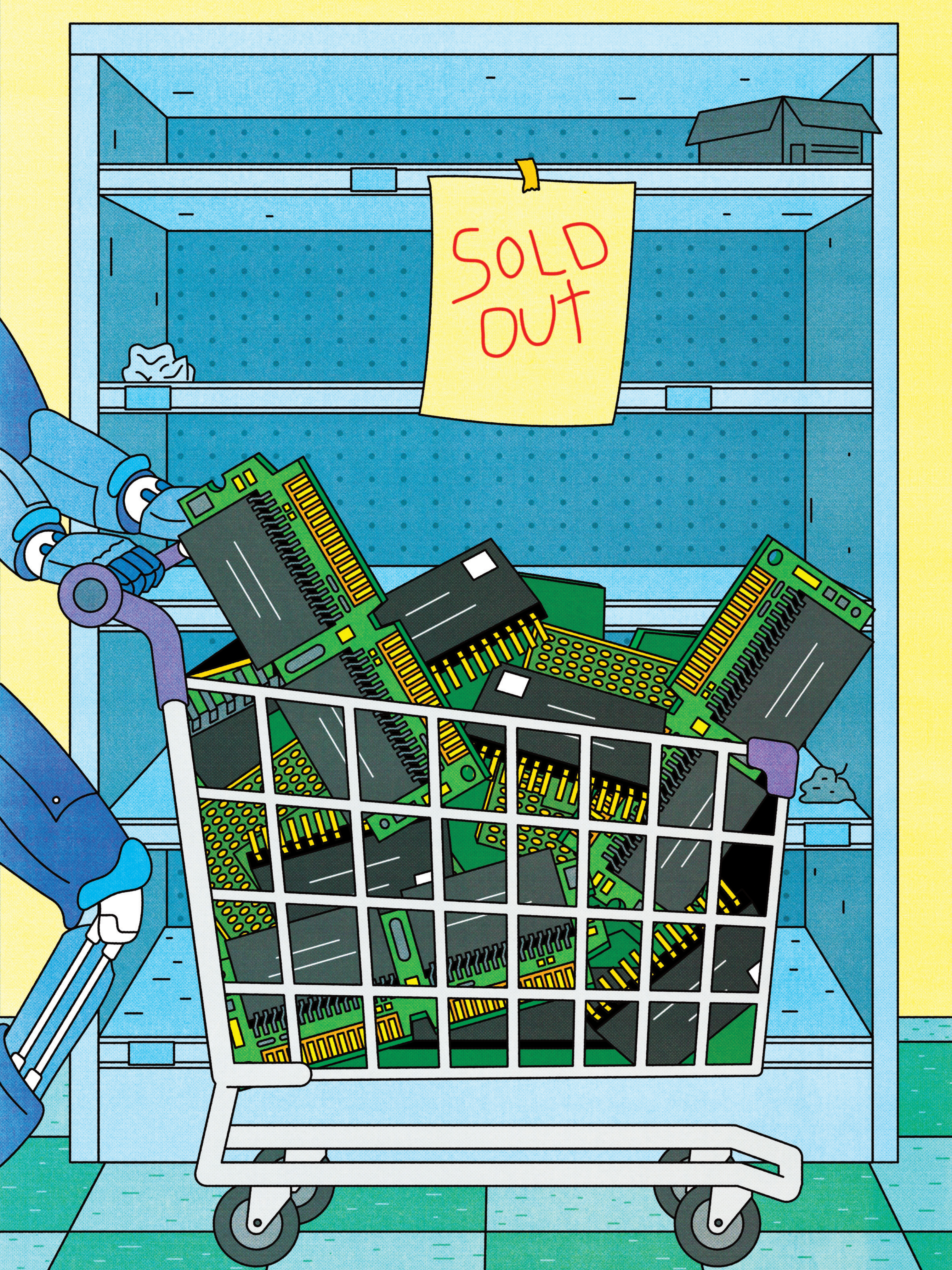
AI is a MEMORY HOG

By SAMUEL K. MOORE
Illustration by George Wylesol

Its demand threatens whole categories of electronics

SEMICONDUCTORS

If it feels these days as if everything in technology is about AI, that's because it is. And nowhere is that more true than in the market for computer memory. Demand, and profitability, for the type of DRAM used to feed GPUs and other accelerators in AI data centers is so huge that it's diverting memory away from the supply needed for other uses and causing prices to skyrocket. According to Counterpoint Research, DRAM prices rose at least 80 to 90 percent in the first quarter of 2026. • The largest AI hardware companies are likely first in line for new memory, but that leaves everybody else—makers of PCs, smartphones, TVs, and everything else that needs to temporarily store billions of bits—scrambling to deal with scarce supply and inflated prices. • How did the electronics industry get into this mess, and even more important, how will it get out? *IEEE Spectrum* asked economists and memory experts to explain. Today's situation, they say, is the result of a collision between the DRAM industry's historic boom-and-bust cycle and an AI hardware infrastructure build-out that's without precedent in its scale. And, barring some major collapse in the AI sector, it will take years for new capacity and new technology to bring supply in line with demand. Even then, prices might stay high. • To understand both the supply and the demand ends of this episode, you need to understand its main culprit: high-bandwidth memory, or HBM.


SOLD OUT

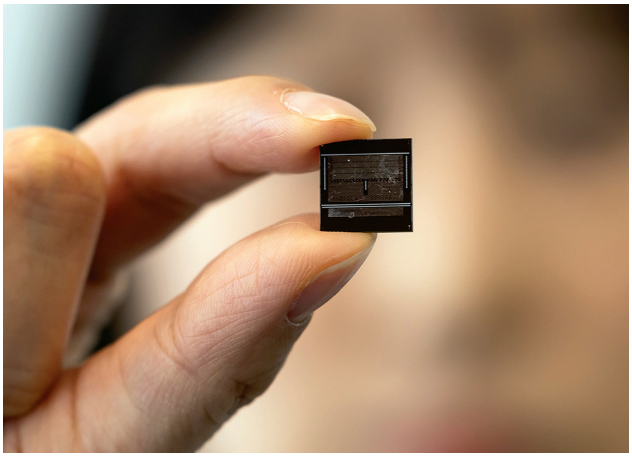
HBM emerged from the DRAM industry's attempt to circumvent the slowing pace of Moore's Law by using 3D chip-packaging technology. Each HBM chip is made up of as many as 12 thinned-down DRAM chips called dies. Each die contains a number of vertical connections called through-silicon vias (TSVs). The dies are piled atop one another and connected by arrays of microscopic solder balls aligned to the TSVs. This DRAM tower—well, at about 750 micrometers thick, it's more of a brutalist office block than a tower—is mounted atop the base die, which shuttles bits between the memory dies and the processor.

This assemblage is then set within a few millimeters of a GPU, to which it is linked by as many as 2,048 micrometer-scale connections. HBMs are attached on two sides of the GPU, and they are all packaged together as a single unit.

The idea behind such a tight, highly parallel connection with the GPU is to knock down what's known as the memory wall. That's the barrier, in energy and time, to providing the GPU with the terabytes of data needed every second to run large language models. Memory bandwidth is a key limiter to how fast LLMs can run.

As the size of AI models has grown, so has HBM's importance to the GPU. HBM has been around for more than a decade, although DRAM makers have boosted its speed and capacity considerably in recent years. But that improvement has come at a cost. The consulting firm SemiAnalysis estimates that HBM generally costs three times as much as other types of memory and constitutes half or more of the cost of the packaged GPU.

People who watch the memory and storage industry agree that DRAM is a highly cyclical market with lavish booms and devastating busts. With new fabs now costing US $15 billion or more, firms are extremely reluctant to build new facilities and may only have the cash to do so during boom times, explains Thomas Coughlin, a memory and storage consultant. But building such a fab and getting it up and running can take two years or more, practically ensuring that new capacity arrives well past the initial surge in demand, flooding the market and depressing prices.



This little thing—a Samsung HBM3e memory chip—is the cause of a lot of headaches.

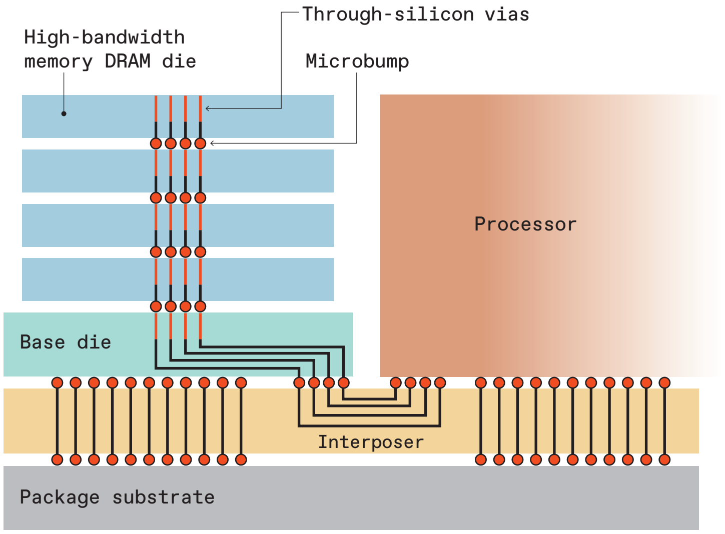
The origins of today's cycle, says Coughlin, go all the way back to the chip-supply panic triggered by the COVID-19 pandemic. To avoid supply-chain stumbles and support the rapid shift to remote work, hyperscalers—data center giants like Amazon, Google, and Microsoft—bought up huge inventories of memory and storage, boosting prices, he notes.

But then supply became more regular and data center expansion fell off in 2022, causing memory and storage prices to plummet. This recession continued into 2023, and even resulted in big memory and storage companies such as Samsung cutting production by 50 percent to try to keep prices from going below the costs of manufacturing, adds Coughlin. It was a rare and fairly desperate move, because companies typically have to run plants at full capacity just to earn back their value.

After a recovery began in late 2023, "all the memory and storage companies were very wary of increasing their production capacity again," says Coughlin. "Thus there was little or no investment in new production capacity in 2024 and through most of 2025."

That lack of new investment is colliding headlong with a huge boost in demand from new data centers. Globally, there are nearly 2,000 new data centers either planned or under construction right now, according to the analysis firm Data Center Map. If they're all built, it would represent about a 20 percent jump in the global supply, which stands at around 9,000 facilities now.

If the build-out continues at its current pace, McKinsey & Co. predicts companies will spend



Inside High-Bandwidth Memory

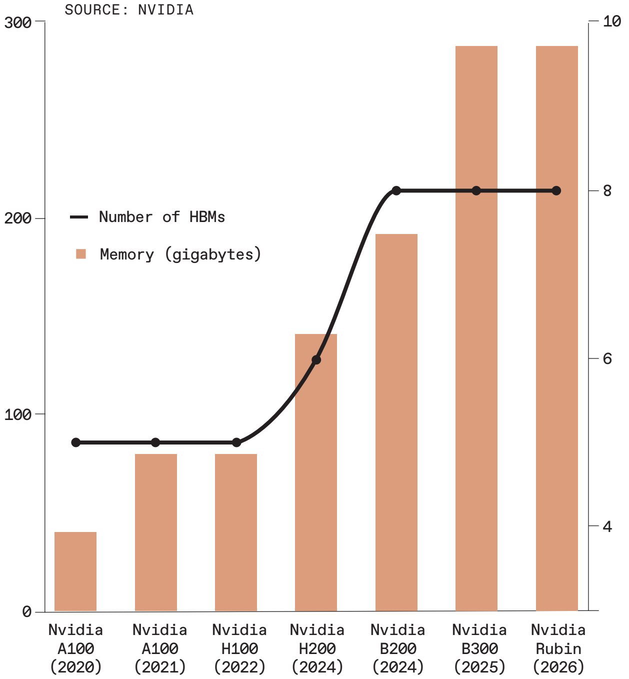
High-bandwidth memory is a stack of DRAM dies on top of a base die. Data flows through the dies by way of vertical connections called through-silicon vias. The HBM sits snugly beside the GPU it serves, and both are attached to a substrate called an interposer.

TOP: SEONGJOON CHO/BLOOMBERG/GETTY IMAGES

GPUs use more and more memory

During the last six years, Nvidia GPUs have used more and more high-bandwidth memory.

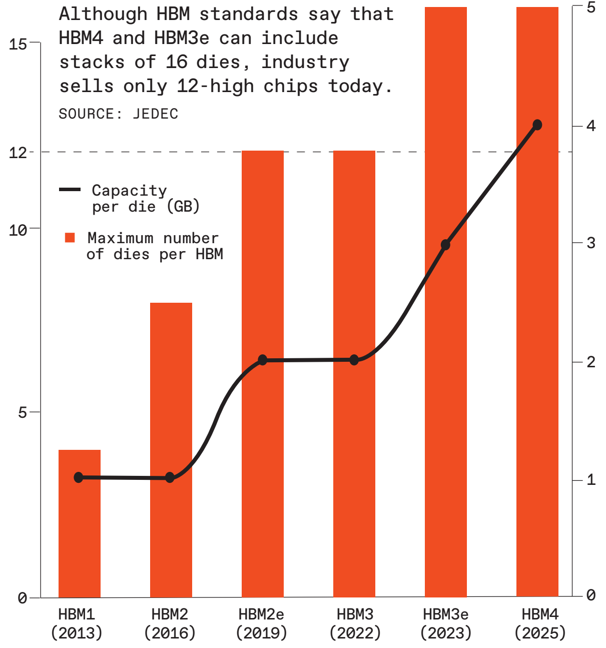
SOURCE: NVIDIA



HBM's progress

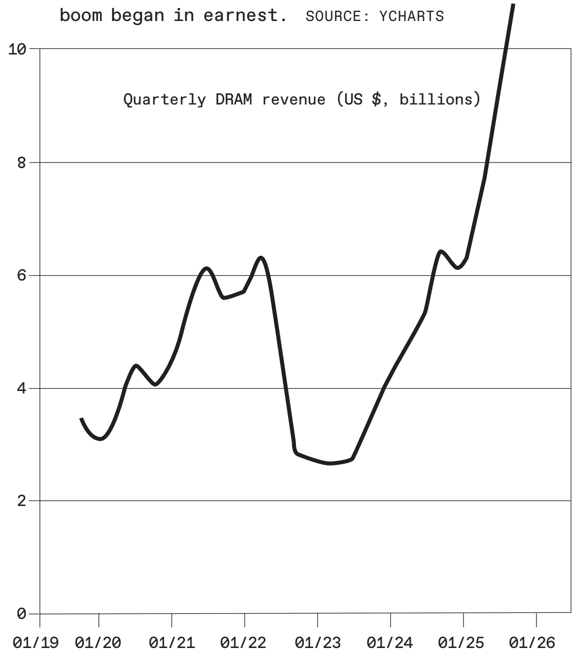
High-bandwidth memory (HBM) standards have called not only for more stacked dies but also more memory capacity per die. Although HBM standards say that HBM4 and HBM3e can include stacks of 16 dies, industry sells only 12-high chips today.

SOURCE: JEDEC



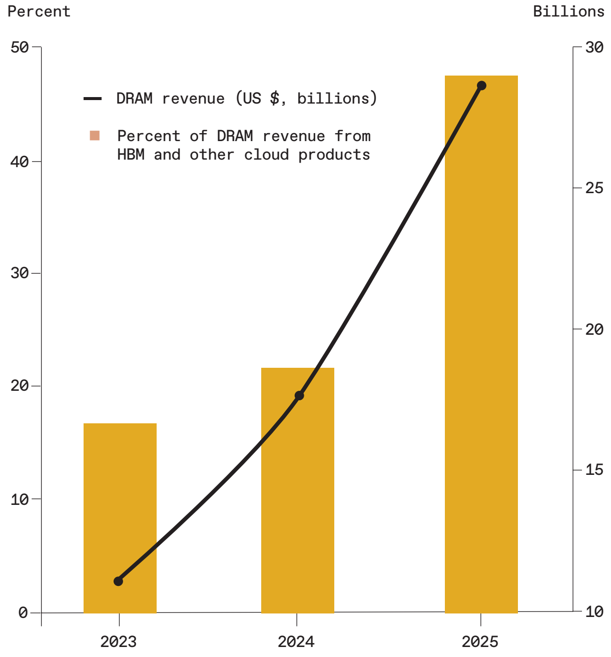
The ups and downs of Micron's DRAM fortunes

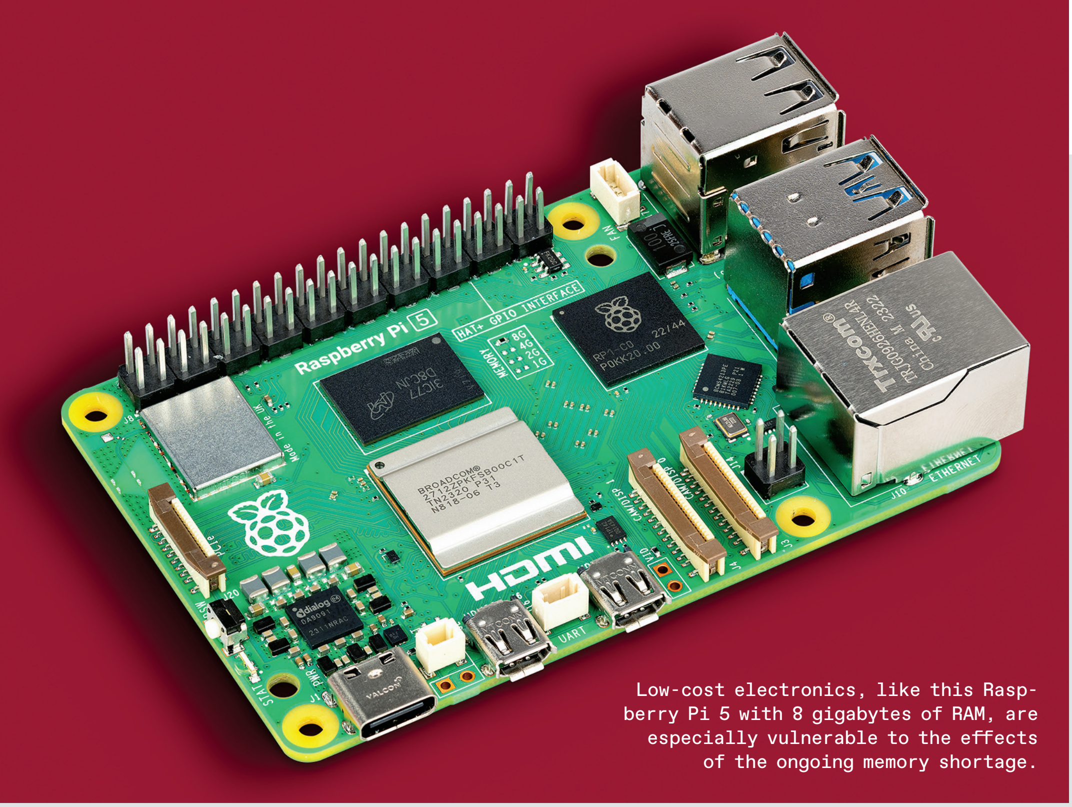
Micron Technology's fortunes are emblematic of the memory industry's boom-and-bust cycles. After a pandemic-era boom, its DRAM revenues crashed and didn't recover until the AI boom began in earnest. SOURCE: YCHARTS



HBM is increasingly valuable

Revenue from high-bandwidth memory and other data center memory products make up an increasing fraction of Micron's DRAM revenue. SOURCE: MICRON





Low-cost electronics, like this Raspberry Pi 5 with 8 gigabytes of RAM, are especially vulnerable to the effects of the ongoing memory shortage.

Memory Shortage Hits Low-Cost Computing Hardest

Surging memory prices are putting all computer makers in a difficult spot, but none more so than low-cost-computer makers. Raspberry Pi Holdings, which produces single-board computers for Internet of Things products and hobbyists, has hiked the price of its Raspberry Pi 5, with 16 gigabytes of RAM, from US $120 in November 2025 to $205 in early March. Framework, a company that makes highly configurable and repairable laptops based on Intel microprocessors, has announced three rounds of memory-related price hikes. Others, such as Orange Pi, have made no official comment, but the price of an Orange Pi 5B with 16 GB of RAM has surged from $160 at the start of 2025 to around $300 in early March.

"If you have a product that's relatively low cost, the memory is going to be a relatively large portion of it," says Raspberry Pi CEO Eben Upton. Most Raspberry Pi computers with the same model number use the same board design and the same hardware components—except for the memory, which can be upgraded to suit the user's needs. With little else to differentiate among its models, Raspberry Pi has to pass increased memory costs on to customers.

The way memory is produced puts many computer makers in a difficult position. "Generally, you have a single fungible pool of manufacturing capacity in DRAM that you can use to do anything," Upton explains. "It can be used to make commodity DRAM, DDR [Double Data Rate], and LPDDR [Low Power DDR], or you can use it to make HBM"—the type most commonly used for AI hardware.

That means low-cost-computer manufacturers are competing with the AI hardware giants for the output of the same pool of manufacturing capacity. And with the world's most valuable tech companies spending billions on AI infrastructure, these manufacturers have little hope of negotiating lower prices.

Larger computer manufacturers can mitigate the memory price shock by working out larger or longer contracts in exchange for lower prices, or by tolerating a lower profit margin. Others are rumored to have stocked up on memory as prices surged.

But these strategies aren't available to companies that sell computers at low price points or in low volumes. The profit margin on such devices is too slim to absorb increased memory costs or to stockpile memory. Companies like Raspberry Pi also purchase in low volumes compared with mainstream PC makers like Lenovo, which means it's difficult to negotiate a volume discount.

IT'S A PERFECT STORM for low-cost computing and one that, in contrast to 2025's U.S. tariff hikes, has led to immediate and unavoidable problems. While tariffs did place some pressure on price for low-cost computers, that pressure wasn't uniformly felt. Raspberry Pi, which manufactures its computers in the United Kingdom, found itself in a better position than competitors who manufactured in China. The memory price increase, on the other hand, applies to all companies in this industry, no matter where or how manufacturing takes place.

Companies that build PCs with removable and user-upgradeable memory are turning toward a bring-your-own-memory approach. Framework offered this option before memory prices increased, but several specialty-desktop manufacturers have recently announced similar bring-your-own-RAM options. However, this isn't possible for many low-cost computers, including those from Raspberry Pi, because they solder the memory to the mainboard.

So Raspberry Pi has adopted a different strategy. A new iteration of the Pi 4 Model B moved from a single memory module to a dual-module configuration. "As you can generally buy smaller RAM more easily than larger RAM, we can use a pair of back-to-back modules instead of a single larger one. You have more vendor diversity, more vendor flexibility," Upton says. He expects the current pricing will remain the same, but the change gives more options when looking to source memory in the future.

RASPBERRY PI

Of course, there’s another strategy all low-cost-computing companies can use. They can simply offer less memory.

Raspberry Pi introduced a version of the Pi 5 with 1 GB of RAM in December 2025, one-sixteenth of the maximum capacity. It debuted at $45 and was the only Pi 5 model to avoid the February price increase. Raspberry Pi was among the first to reduce system memory, and in retrospect the decision looks prescient.

And the problem is starting to leak through to mainstream consumer electronics. Samsung’s new Galaxy S26 is $100 more than the prior year’s model; the company cited memory pricing as a factor. Apple also increased the price of its newest MacBook Air by $100 and stuck to just 8 GB of memory for its low-cost MacBook Neo. All other Macs ship with at least 16 GB. Analysts expect more price increases through 2026.

So when will memory prices come down—or at least stop rising?

Upton expects the timing to be similar to past memory price cycles, which means it’s likely to last for at least a few years. “Memory will be expensive this year. It will probably be expensive next year,” he says, adding that he’d be “a little bit surprised” if price increases have not leveled off into 2028. In any case, he cautions against being too sure of the future. During the COVID-19 pandemic chip shortage, companies and consumers worried that inexpensive logic chips were a thing of the past, but the situation eventually returned to normal. “Like all bubbly phenomena, it’s very hard to measure.”

–MATTHEW S. SMITH

$7 trillion by 2030, with the bulk of that—$5.2 trillion—going to AI-focused data centers. Of that chunk, $3.3 trillion will go toward servers, data storage, and network equipment, the firm predicts.

The biggest beneficiary so far of this boom is unquestionably the GPU maker Nvidia. Revenue for its data center business went from barely a billion in the final quarter of 2019 to $51 billion in the quarter that ended in October 2025. Over this period, each of Nvidia’s server GPUs has demanded an increasing number of DRAM chips. The recently released B300 uses eight HBM chips, each of which is a stack of 12 DRAM dies. Competitors’ use of HBM has largely mirrored Nvidia’s. AMD’s MI350 GPU, for example, also uses eight 12-die chips.

With so much demand for AI infrastructure, an increasing fraction of the revenue for DRAM makers comes from HBM. Micron Technology—the No. 3 producer behind SK Hynix and Samsung—reported that HBM and other cloud-related memory went from being 17 percent of its DRAM revenue in 2023 to nearly 50 percent in 2025.

Micron predicts the total market for HBM will grow from $35 billion in 2025 to $100 billion by 2028—a figure larger than the entire DRAM market was in 2024, CEO Sanjay Mehrotra told analysts in December. It’s reaching that figure two years earlier than Micron had previously expected. Across the industry, demand will outstrip supply “substantially… for the foreseeable future,” he said.

T **“There are two ways** to address supply issues with DRAM: with innovation or with building more fabs,” explains Mina Kim, an economist with Mkecon Insights. “As DRAM scaling has become more difficult, the industry has turned to advanced packaging…which is just using more DRAM.”

Micron, Samsung, and SK Hynix combined make up the vast majority of the memory market, and all three have new fabs and facilities in the works. However, these are unlikely to help bring down prices substantially in the next couple of years.

- Micron is in the process of building an HBM fab in Singapore that should be in production in 2027. And it is retooling a fab it purchased from Taiwan-based Powerchip Semiconductor Manufacturing Co. that will begin production in the second half of 2027. In January, Micron broke ground on what will be a DRAM fab complex in Onondaga County, N.Y. It won’t be in full production until 2030.
- Samsung plans to start producing at a new plant in Pyeongtaek, South Korea, in 2028.
- SK Hynix is building HBM and packaging facilities in West Lafayette, Ind., that are set to begin production by the end of 2028. An HBM fab the company’s building in Cheongju, South Korea, should be complete in 2027.

With these expansions unable to contribute for several years, other factors will be needed to increase supply. “Relief will come from a combination of incremental capacity expansions by existing DRAM leaders, yield improvements in advanced packaging, and a broader diversification of supply chains,” says Shawn DuBravac, chief economist for the Global Electronics Association (formerly the Institute of Printed Circuits). “New fabs will help at the margin, but the faster gains will come from process learning, better [DRAM] stacking efficiency, and tighter coordination between memory suppliers and AI chip designers.”

So, will prices come down once some of these new plants come on line? Don’t bet on it. “In general, economists find that prices come down much more slowly and reluctantly than they go up. DRAM today is unlikely to be an exception to this general observation, especially given the insatiable demand for compute,” says Kim.

In the meantime, technologies are in the works that could make HBM an even bigger consumer of silicon. The standard for HBM4 can accommodate 16 stacked DRAM dies, even though today’s chips use only 12 dies. Getting to 16 has a lot to do with the chip-stacking technology. Conducting heat through the HBM “layer cake” of silicon, solder, and support material is a key limiter to going higher and in repositioning HBM inside the package to get even more bandwidth.

SK Hynix claims a heat-conduction advantage through a manufacturing process called advanced MR-MUF (mass reflow-molded underfill). Further out, an alternative chip-stacking technology called hybrid bonding could help heat conduction by reducing the die-to-die vertical distance essentially to zero. In 2024, researchers at Samsung proved they could produce a 16-high stack with hybrid bonding, and they suggested that 20 dies was not out of reach.

All that new technology and new capacity could arrive at an inopportune time, if AI’s star starts to fade. “Not to be too cynical, but the current investments in AI expansion can’t last,” says Coughlin. Depending on the timing, “we may see another of the boom-and-bust cycles for this industry as demand drops at the same time as supply increases.” ■

A Smart Pill That Can Sense, Sample, and Treat

BY REZA GHODSSI,
JUSTIN STINE
& LUKE BEARDSLEE

Ingestible electronics will deliver drugs and take biopsies

Photo-illustration by Edmon de Haro

One day soon, a doctor might prescribe a pill that doesn't just deliver medicine but also reports back on what it finds inside you—and then takes actions based on its findings.

Instead of scheduling an endoscopy or CT scan, you'd swallow an electronic capsule smaller than a multivitamin. As it travels through your digestive system, it could check tissue health, look for cancerous changes, and send data to your doctor. It could even release drugs exactly where they're needed or snip a tiny biopsy sample before passing harmlessly out of your body.

This dream of a do-it-all pill is driving a surge of research into ingestible electronics: smart capsules designed to monitor and even treat disease from inside the gastrointestinal (GI) tract. The stakes are high. GI diseases affect tens of millions of people worldwide, including such ailments as inflammatory bowel disease, celiac disease, and small intestinal bacterial overgrowth. Diagnosis often involves a frustrating maze of blood tests, imaging, and invasive endoscopy. Treatments, meanwhile, can bring serious side effects because drugs affect the whole body, not just the troubled gut.

If capsules could handle much of that work—streamlining diagnosis, delivering targeted therapies, and sparing patients repeated invasive procedures—they could transform care. Over the past 20 years, researchers have built a growing tool kit of ingestible devices, some already in clinical use. These capsule-shaped devices typically contain sensors, circuitry, a power source, and sometimes a communication module, all enclosed in a biocompatible shell. But the next leap forward is still in development: autonomous capsules that can both sense and act, releasing a drug or taking a tissue sample.

That's the challenge that our lab—the MEMS Sensors and Actuators Laboratory (MSAL) at the University of Maryland, College Park—is tackling. Drawing on decades of advances in microelectromechanical systems (MEMS), we're building swallowable devices that integrate sensors, actuators, and wireless links in packages that are small and safe enough for patients. The hurdles are considerable: power, miniaturization, biocompatibility, and reliability, to name a few. But the potential payoff will be a new era of personalized and minimally invasive medicine, delivered by something as simple as a pill you can swallow at home.

The idea of a smart capsule has been around since the late 1950s, when researchers first experimented with swallowable devices to record temperature, gastric pH, or pressure inside the digestive tract. At the time, it seemed closer to science fiction than clinical reality, bolstered by pop-culture visions like the 1966 film *Fantastic Voyage*, where miniaturized doctors travel inside the human body to treat a blood clot.

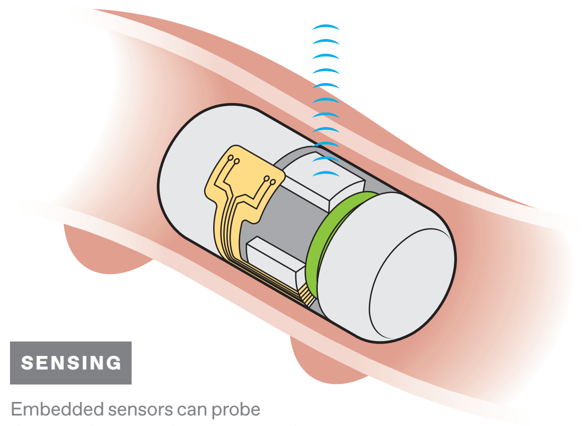
For decades, though, the mainstay of GI diagnostics was endoscopy: a camera on a flexible tube, threaded down the throat or up through the colon. These procedures are quite invasive and require patients to be sedated, which increases both the risk of complications and procedural costs. What's more, it's difficult for endoscopes to safely traverse the circuitous pathway of the small intestine. The situation changed in the early 2000s, when video-capsule endoscopy arrived. The best-known product, PillCam, looks like a large vitamin but contains a camera, LEDs, and a transmitter. As it passes through the gut, it beams images and videos to a wearable device.

Today, capsule endoscopy is a routine tool in gastroenterology; ingestible devices can measure acidity, temperature, or gas concentrations. And researchers are pushing further, with experimental prototypes that deliver drugs or analyze the microbiome. For example, teams from Tufts University, in Massachusetts, and Purdue University, in Indiana, are working on devices with dissolvable coatings and mechanisms to collect samples of liquid for studies of the intestinal microbiome.

Still, all those devices are passive. They activate on a timer or by exposure to the neutral pH of the intestines, but they

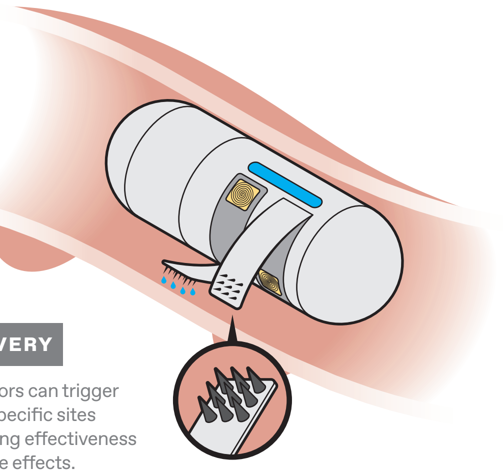
WHAT SMART CAPSULES CAN DO

Ingestible electronic capsules use miniaturized sensors and actuators to monitor the gut, deliver medication, and collect biological samples.



SENSING

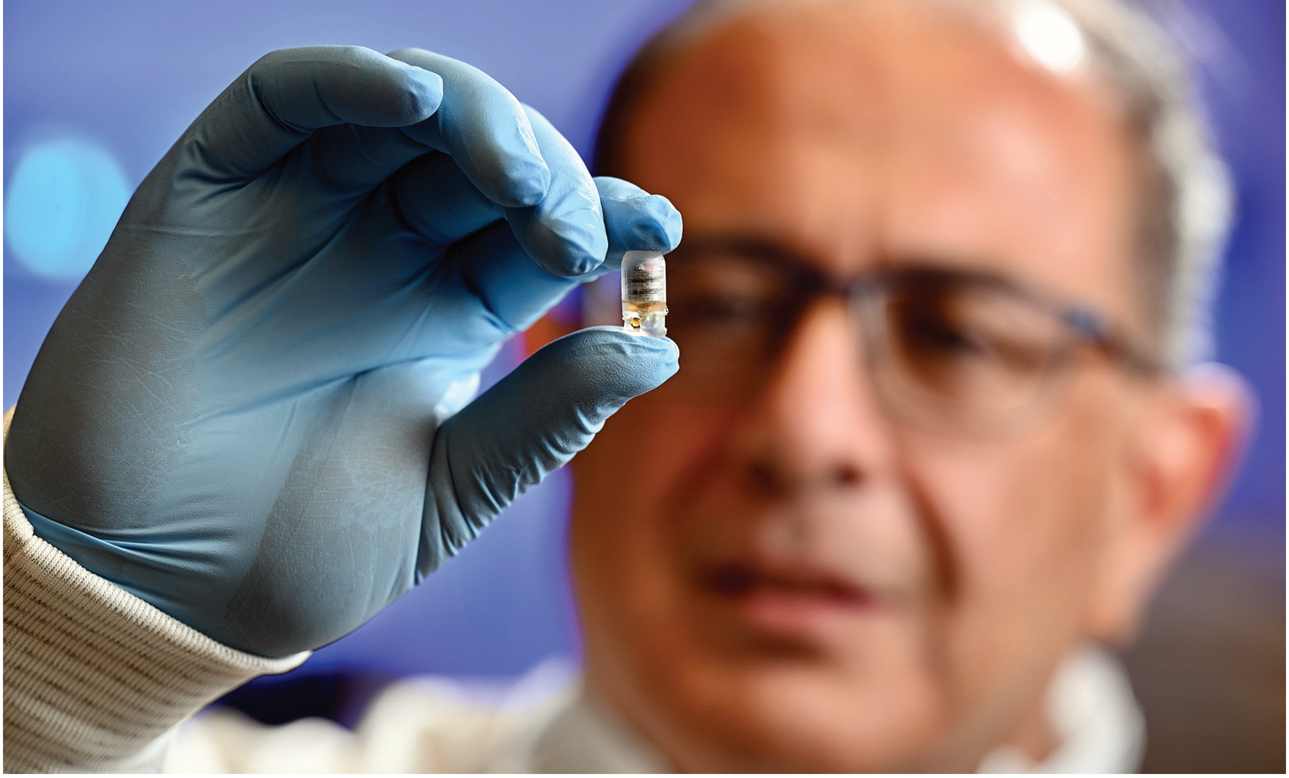
Embedded sensors can probe the gut—for example, measuring the bioimpedance of the intestinal lining to detect disease—and transmit the data wirelessly.



DRUG DELIVERY

Miniature actuators can trigger drug release at specific sites in the gut, boosting effectiveness while limiting side effects.

Illustrations by Chris Philpot

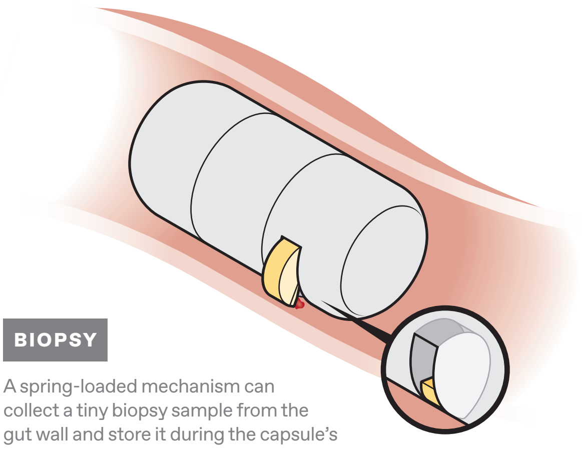


One of the authors (Ghodssi) holds a miniaturized drug-delivery capsule that's designed to release medication at specific sites in the gastrointestinal tract.

MAXIMILIAN FRANZ/ENGINEERING AT MARYLAND MAGAZINE

don't adapt to conditions in real time. The next step requires capsules that can sense biomarkers, make decisions, and trigger specific actions—moving from clever hardware to truly autonomous "smart pills." That's where our work comes in.

Since 2017, MSAL has been pushing ingestible devices forward with the goal of making an immediate impact in health care. The group built on the MEMS community's legacy in microfabrication, sensors, and system integration, while taking advantage of new tools like 3D printing and materials like biocompatible polymers. Those advances have made it possible to prototype faster and shrink devices smaller, sparking a wave of innovation in wearables, implants, and now ingestibles. Today, MSAL is collaborating with engineers, physicians, and data scientists to move these capsules from lab benches to pharmaceutical trials.



BIOPSY

A spring-loaded mechanism can collect a tiny biopsy sample from the gut wall and store it during the capsule's passage through the digestive system.

As a first step, back in 2017, we set out to design sensor-carrying capsules that could reliably reach the small intestine and indicate when they reached it. Another challenge was that sensors that work well on the benchtop can falter inside the gut, where shifting pH, moisture, digestive enzymes, and low-oxygen conditions can degrade typical sensing components.

Our earliest prototype adapted MEMS sensing technology to detect abnormal enzyme levels in the duodenum that are linked to pancreatic function. The sensor and its associated electronics were enclosed in a biocompatible, 3D-printed shell coated with polymers that dissolved only at certain pH levels. This strategy could one day be used to detect biomarkers in secretions from the pancreas to detect early-stage cancer.

That first effort with a passive device taught us the fundamentals of capsule design and opened the door to new applications. Since then, we've developed sensors that can track biomarkers such as the gas hydrogen sulfide, neurotransmitters such as serotonin and dopamine, and bioimpedance—a measure of how easily ions pass through intestinal tissue—to shed light on the gut microbiome, inflammation, and disease progression. In parallel, we've worked on more-active devices: capsule-based tools for controlled drug release and tissue biopsy, using low-power actuators to trigger precise mechanical movements inside the gut.

Like all new medical devices and treatments, ingestible electronics face many hurdles before they reach patients—from earning physician trust and insurance approval to demonstrating clear benefits, safety, and reliability. Packaging is a particular focus, as the capsules must be easy to swallow yet durable enough to survive stomach acid. The field is steadily proving safety and reliability, progressing from proof of concept in tissue, through the different stages of animal studies, and eventually to human trials. Every stage provides evidence that reassures doctors and patients—for example, showing that ingesting a properly packaged tiny battery is safe, and that a capsule's wireless signals, far weaker than those of a cellphone, pose no health risk as they pass through the gut.

The gastrointestinal tract is packed with clues about health and disease, but much of it remains out of reach of standard diagnostic tools. Ingestible capsules offer a way in, providing direct access to the small intestine and colon. Yet in many cases, the concentrations of chemical biomarkers can be too low to detect reliably in early stages of a disease, which makes the engineering challenge formidable. What's more, the gut's corrosive, enzyme-rich environment can foul sensors in multiple ways, interfering with measurements and adding noise to the data.

Take, for example, inflammatory bowel disease, for which there is no standard clinical test. Rather than searching for a scarce biomarker molecule, our team focused on a physical change: the permeability of the gut lining, which is a key factor in the disease. We designed capsules that measure the intestinal tissue's bioimpedance by sending tiny currents across electrodes and recording how the tissue resists or conducts those currents at different frequencies (a technique called impedance spectroscopy). To make the electrodes suitable for in vivo use, we coated them with a thin, conductive, biocompatible polymer that reduces electrical noise and keeps stable contact with the gut wall. The capsule finishes its job by transmitting its data wirelessly to our computers.

In our lab tests, the capsule performed impressively, delivering clean impedance readouts from excised pig tissue even when the sample was in motion. In our animal studies, it detected shifts in permeability triggered by calcium chelators, compounds that pry open the tight junctions between intestinal cells. These results suggest that ingestible bioimpedance capsules could one day give clinicians a direct, minimally invasive window into gut-barrier function and inflammation. We believe that ingestible diagnostics can serve as powerful tools—catching disease earlier, confirming whether treatments are working, and establishing a baseline for gut health.

Targeted drug delivery is one of the most compelling applications for ingestible capsules. Many drugs for GI conditions—such as biologics for inflammatory bowel disease—can cause serious side effects that limit both dosage and duration of treatment. A promising alternative is delivering a drug directly to the diseased tissue. This localized approach boosts the drug's concentration at the target site while reducing its spread throughout the body, which improves effectiveness and minimizes side effects. The challenge is engineering a device that can both recognize diseased tissue and deliver medication quickly and precisely.

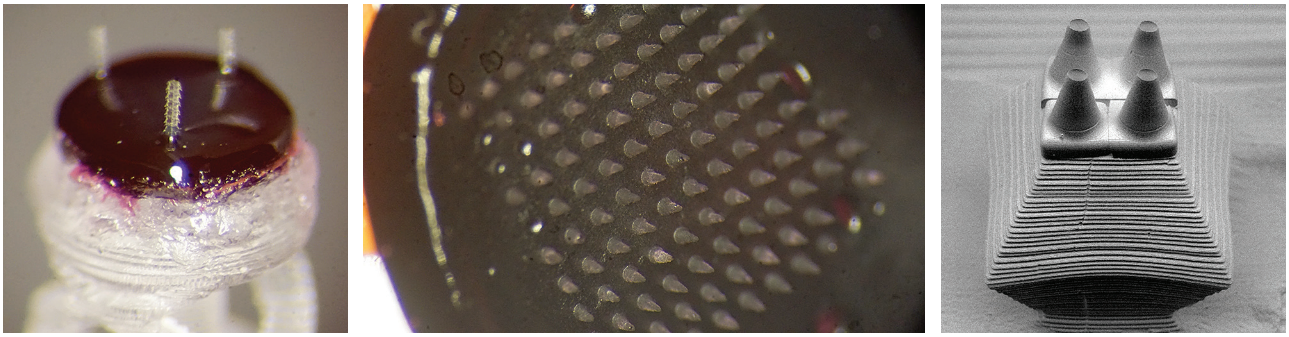
With other labs making great progress on the sensing side, we've devoted our energy to designing devices that can deliver the medicine. We've developed miniature actuators—tiny moving parts—that meet strict criteria for use inside the body: low power, small size, biocompatibility, and long shelf life.

Some of our designs use soft and flexible polymer "cantilevers" with attached microneedle systems that pop out from the capsule with enough force to release a drug, but without harming the intestinal tissue. While hollow microneedles can directly inject drugs into the intestinal lining, we've also demonstrated prototypes that use the microneedles for anchoring drug payloads, allowing the capsule to release a larger dose of medication that dissolves at an exact location over time.

In other experimental designs, we had the microneedles themselves dissolve after injecting a drug. In still others, we used microscale 3D printing to tailor the structure of the microneedles and control how quickly a drug is released—providing either a slow and sustained dose or a fast delivery. With this 3D printing, we created rigid microneedles that penetrate the mucosal lining and gradually diffuse the drug into the tissue, and soft microneedles that compress when the cantilever pushes them against the tissue, forcing the drug out all at once.

Tissue sampling remains the gold standard diagnostic tool in gastroenterology, offering insights far beyond what doctors can glean from visual inspection or blood tests. Capsules hold unique promise here: They can travel the full length of the GI tract, potentially enabling more frequent and affordable biopsies than traditional procedures. But the engineering hurdles are substantial. To collect a sample, a device must generate significant mechanical force to cut through the tough, elastic muscle of the intestines—while staying small enough to swallow.

Microneedle designs for drug-delivery capsules have evolved over the years. An early prototype [left] used microneedle anchors to hold a capsule in place. Later designs adopted molded microneedle arrays [center] for more uniform fabrication. The most recent version [right] integrates hollow microinjector needles, allowing more precise and controllable drug delivery.



FROM LEFT: UNIVERSITY OF MARYLAND/WILEY; UNIVERSITY OF MARYLAND/ELSEVIER; UNIVERSITY OF MARYLAND/ACS

Different strategies have been explored to solve this problem. Torsion springs can store large amounts of energy but are difficult to fit inside a tiny capsule. Electrically driven mechanisms may demand more power than current capsule batteries can provide. Magnetic actuation is another option, but it requires bulky external equipment and precise tracking of the capsule inside the body.

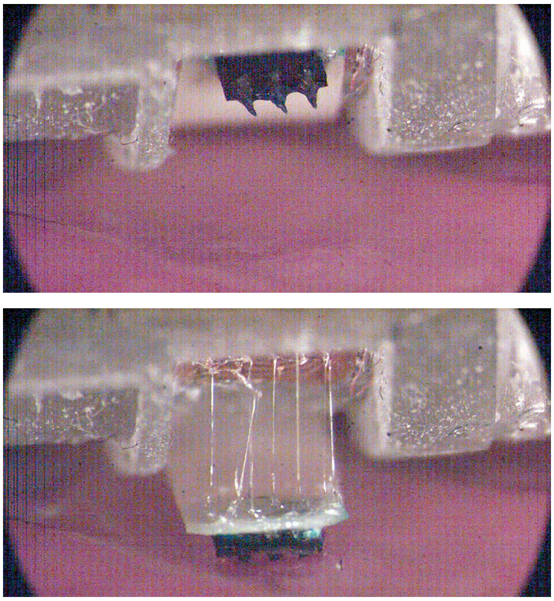
Our group has developed a low-power biopsy system that builds on the torsion-spring approach. We compress a spring and use adhesive to "latch" it closed within the capsule, then attach a microheater to the latch. When we wirelessly send current to the device, the microheater melts the adhesive on the latch, triggering the spring. We've experimented with tissue-collection tools, integrating a bladed scraper or a biopsy punch (a cylindrical cutting tool) with our spring-activated mechanisms; either of those tools can cut and collect tissue from the intestinal lining. With advanced 3D printing methods like direct laser writing, we can put fine, microscale edges on these miniature cutting tools that make it easier for them to penetrate the intestinal lining.

Storing and protecting the sample until the capsule naturally passes through the body is a major challenge, requiring both preservation of the sample and resealing the capsule to prevent contamination. In one of our designs, residual tension in the spring keeps the bladed scraper rotating, pulling the sample into the capsule and effectively closing a hatch that seals it inside.

Looking ahead, we expect to see the first clinical applications emerge in early-stage screening. Capsules that can detect electrochemical, bioimpedance, or visual signals could help doctors make sense of symptoms like vague abdominal pain by revealing inflammation, gut permeability, tumors, or bacterial overgrowth. They could also be adapted to screen for GI cancers. This need is pressing: The American Cancer Society reports that as of 2021, 41 percent of eligible U.S. adults were not up to date on colorectal cancer screening. What's more, effective screening tools don't yet exist for some diseases, such as small bowel adenocarcinoma. Capsule technology could make screening less invasive and more accessible.

Of course, ingestible capsules carry risks. The standard hazards of endoscopy still apply, such as the possibility of bleeding and perforation, and capsules introduce new complications. For example, if a capsule gets stuck in its passage through the GI tract, it could cause bowel obstruction and require endoscopic retrieval or even surgery. And concerns that are specific to ingestibles, including the biocompatibility of materials, reliable encapsulation of electronics, and safe battery operation, all demand rigorous testing before clinical use.

Powering these capsules is a key challenge that must be solved on the path to the clinic. Most capsule endoscopes today rely on coin-cell batteries, typically silver oxide, which offer a safe and energy-dense source but often occupy 30 to 50 percent of the capsule's volume. So researchers have investigated alternatives, from wireless power transfer to energy-harvesting systems. At the State University of New York at Binghamton, one team is exploring microbial fuel cells that generate electricity from probiotic bacteria interacting with nutrients in the gut. At MIT, researchers used the gastric fluids of a pig's stomach to power a simple battery. In our own lab, we are exploring piezoelectric and electrochemical approaches to harvesting energy throughout the GI tract.


Stills from a high-speed video show how a capsule deploys microneedles to deliver drugs into intestinal tissue.

UNIVERSITY OF MARYLAND/ELSEVIER

The next steps for our team are pragmatic ones: working with gastroenterologists and animal-science experts to put capsule prototypes through rigorous in vivo studies, then refining them for real-world use. That means shrinking the electronics, cutting power consumption, and integrating multiple functions into a single multimodal device that can sense, sample, and deliver treatments in one pass. Ultimately, any candidate capsule will require regulatory approval for clinical use, which in turn demands rigorous proof of safety and clinical effectiveness for a specific medical application.

The broader vision is transformative. Swallowable capsules could bring diagnostics and treatment out of the hospital and into patients' homes. Whereas procedures with endoscopes require anesthesia, patients could take ingestible electronics easily and routinely. Consider, for example, patients with inflammatory bowel disease who live with an elevated risk of cancer; a smart capsule could perform yearly cancer checks, while also delivering medication directly wherever necessary.

Over time, we expect these systems to evolve into semi-autonomous tools: identifying lesions, performing targeted biopsies, and perhaps even analyzing samples and applying treatment in place. Achieving that vision will require advances at the very edge of microelectronics, materials science, and biomedical engineering, bringing together capabilities that once seemed impossible to combine in something the size of a pill. These devices hint at a future in which the boundary between biology and technology dissolves, and where miniature machines travel inside the body to heal us from within. ■

Rise of the AUTONOMOUS Attack Drones

By **TEREZA PULTAROVA**

1

2

A Russian Geran-2 drone, based on the Iranian Shahed-136, flies over Kyiv during an attack on 27 December 2025 [4]. The Skynode X system, from Auterion, provides a degree of autonomy to a drone [5]. An interceptor drone, part of the U.S. MEROPS defensive system, is tested in Poland on 18 November 2025 [6].

FROM LEFT: SERGEI SUPINSKY/AFP/GETTY IMAGES; AUTERION; WOJTEK RADWANSKI/AFP/GETTY IMAGES

4

5

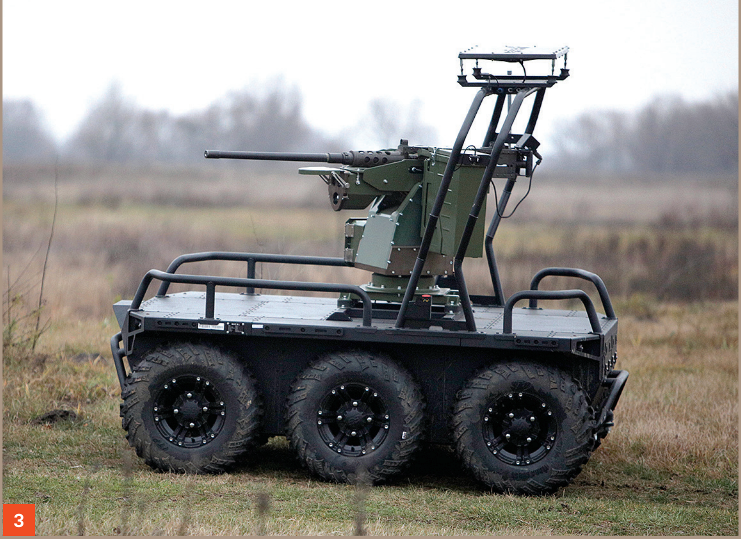
7

8

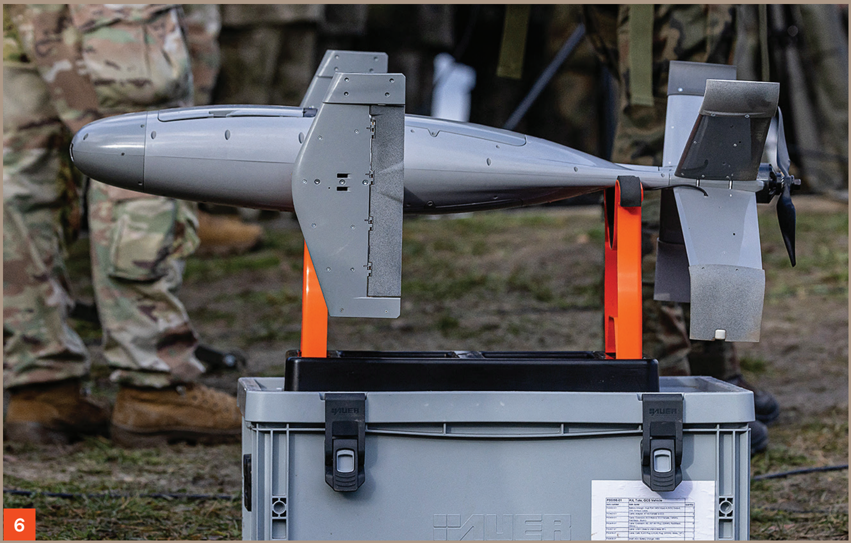
AI-enabled drones being developed by Ukraine and Russia are remaking the battlefield

A fixed-wing drone is tested in Ukraine in April 2025 [1]. Russian V2U attack drones are outfitted with Nvidia processors and run computer-vision software and AI algorithms to enable the drones to navigate autonomously [2]. A Ukrainian land robot, the Ravlyk, can be outfitted with a machine gun [3].

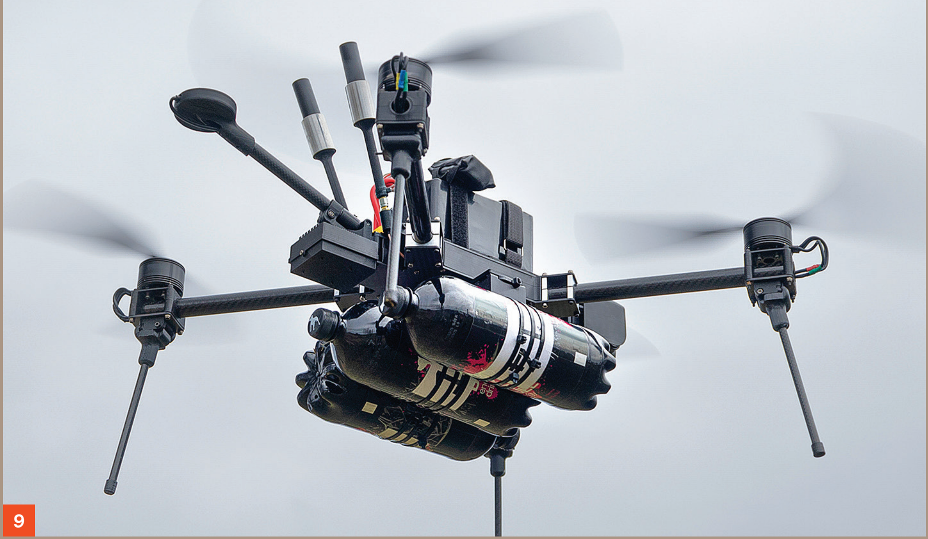
FROM LEFT: ANDREW KRAVCHENKO/BLOOMBERG/GETTY IMAGES; GUR OF THE MINISTRY OF DEFENSE OF UKRAINE; OLEKSANDR KLYMENKO/UKRINFORM/NURPHOTO/GETTY IMAGES



3



6



9

A truck with Stinger-missile launchers, part of a U.S.-developed air-defense system called Avenger, is parked outside Kyiv on 29 October 2025 [7]. During a test in Ukraine, a technician launches a Norda Dynamics Dart-2 fixed-wing strike drone [8]. An experimental drone is demonstrated at the Brave1 defense-tech incubator in Kyiv [9].

FROM LEFT: ANDRIY DUBCHAK/FRONTLINER/GETTY IMAGES; FINBARR O'REILLY; DANYLO DUBCHAK/FRONTLINER/GETTY IMAGES

WHEN KYIV-BORN ENGINEER Yaroslav Azhnyuk thinks about the future, his mind conjures up dystopian images. He talks about "swarms of autonomous drones carrying other autonomous drones to protect them against autonomous drones, which are trying to intercept them, controlled by AI agents overseen by a human general somewhere." He also imagines flotillas of autonomous submarines, each carrying hundreds of drones, suddenly emerging off the coast of California or Great Britain and discharging their cargoes en masse to the sky.

"How do you protect from that?" he asks as we speak in late December 2025; me at my quiet home office in London, he in Kyiv, which is bracing for another wave of missile attacks.

Azhnyuk is not an alarmist. He cofounded and was formerly CEO of Petcube, a California-based company that uses smart cameras and an app to let pet owners keep an eye on their beloved creatures left alone at home. A self-described "liberal guy who didn't even receive military training," Azhnyuk changed his mind about developing military tech in the months following the Russian invasion of Ukraine in February 2022. By 2023, he had relinquished his CEO role at Petcube to do what many Ukrainian technologists have done—to help defend his country against a mightier aggressor.

It took a while for him to figure out what, exactly, he should be doing. He didn't join the military, but through friends on the front line, he witnessed how, out of desperation, Ukrainian troops turned to off-the-shelf consumer drones to make up for their country's lack of artillery.

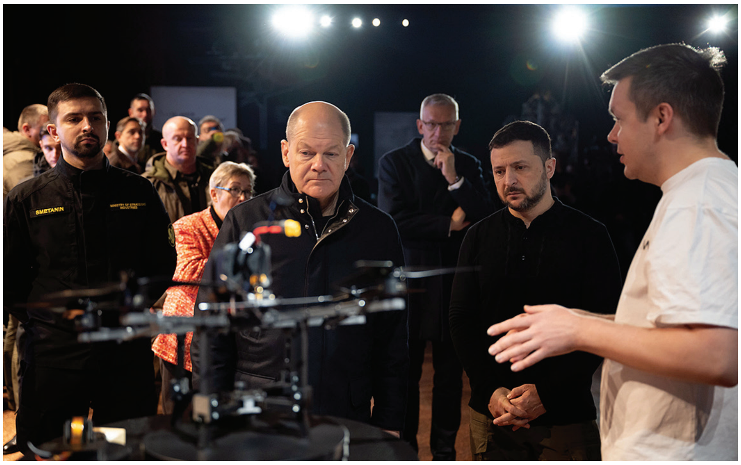
Ukrainian troops first began using drones for battlefield surveillance, but within a few months they figured out how to strap explosives onto them and turn them into effective, low-cost killing machines. Little did they know they were fomenting a revolution in warfare.

That revolution was on display last month, as the U.S. and Israel went to war with Iran. It soon became clear that attack drones are being extensively used by both sides. Iran, for example, is relying heavily on the Shahed drones that the country invented and that are now also being manufactured in Russia and launched by the thousands every month against Ukraine.

A thorough analysis of the Middle East conflict will take some time to emerge. And so to understand the direction of this new way of war, look to Ukraine, where its next phase—autonomy—is already starting to come into view. Outnumbered by the Russians and facing increasingly sophisticated jamming and spoofing aimed at causing the drones to veer off course or fall out of the sky, Ukrainian technologists realized as early as 2023 that what could really win the war was autonomy. Autonomous operation means a drone isn't being flown by a remote pilot, and therefore there's no communications link to that pilot that can be severed or spoofed, rendering the drone useless.

By late 2023, Azhnyuk set out to help make that vision a reality. He founded two companies, The Fourth Law and Odd Systems, the first to develop AI algorithms to help drones overcome jamming during final approach, the second to build thermal cameras to help those drones better sense their surroundings.

"I moved from making devices that throw treats to dogs to making devices that throw explosives on Russian occupants," Azhnyuk quips.



The Ukrainian robotics company The Fourth Law produces an autonomy module [above] that uses optics and AI to guide a drone to its target. Yaroslav Azhnyuk [left, in light shirt], founder and CEO of The Fourth Law, describes a developmental drone with autonomous capabilities to Ukrainian President Volodymyr Zelenskyy and German Chancellor Olaf Scholz.

LEFT: THE PRESIDENTIAL OFFICE OF UKRAINE; RIGHT: THE FOURTH LAW

KOSTYANTYN LIBEROV/LIBKOS/GETTY IMAGES

In the Donetsk Region, on 15 August 2025, a Ukrainian soldier hunts for Shaheds and other drones with a thermal-imaging system attached to a ZU-23 23-millimeter antiaircraft gun.

Since then, The Fourth Law has dispatched "more than thousands" of autonomy modules to troops in eastern Ukraine (it declines to give a more specific figure), which can be retrofitted on existing drones to take over navigation during the final approach to the target. Azhnyuk says the autonomy modules, worth around US $50, increase the drone-strike success rate by up to four times that of purely operator-controlled drones.

And that is just the beginning. Azhnyuk is one of thousands of developers, including some who relocated from Western countries, who are applying their skills and other resources to advancing the drone technology that is the defining characteristic of the war in Ukraine. This eclectic group of startups and founders includes Eric Schmidt, the former Google CEO, whose company Swift Beat is churning out autonomous drones and modules for Ukrainian forces. The frenetic pace of tech development is helping a scrappy, innovative underdog hold at bay a much larger and better-equipped foe.

All of this development is careening toward AI-based systems that enable drones to navigate by recognizing features in the terrain, lock on to and chase targets without an operator's guidance, and eventually exchange information with each other through mesh networks, forming self-organizing robotic kamikaze swarms. Such an attack swarm would be commanded by a single operator from a safe distance.

According to some reports, autonomous swarming technology is also being developed for sea drones. Ukraine has had some notable successes with sea drones, which have reportedly destroyed or damaged around a dozen Russian vessels.

For Ukraine, swarming can solve a major problem that puts the nation at a disadvantage against Russia—the lack of personnel. Autonomy is "the single most impactful defense technology of this century," says Azhnyuk. "The moment this happens, you shift from a manpower challenge to a production challenge, which is much more manageable," he adds.

The autonomous warfare future envisioned by Azhnyuk and others is not yet a reality. But Marc Lange, a German defense analyst and business strategist, believes that "an inflection point" is already in view. Beyond it, "things will be so dramatically different," he says.

"Ukraine pretty rapidly realized that if the operator-to-drone ratio can be shifted from one-to-one to one-to-many, that creates great economies of scale and an amazing cost exchange ratio," Lange adds. "The moment one operator can launch 100, 50, or even just 20 drones at once, this completely changes the economics of the war."

The widespread use of first-person-view (FPV) drones began around early 2023 in Ukraine, changing the dynamics of that war. Traveling at speeds of up to 200 kilometers per hour, these drones transmit video into FPV goggles that show the human operator what the drone sees. There are two main types. FPV killer drones deal sudden death from the sky, while FPV reconnaissance drones, nimble and inconspicuous, sneak overhead with a quiet buzz, making the battlefield completely transparent. The internet is full of chilling video segments, filmed by FPV drones, showing the helpless final seconds of hunted soldiers.

"Anything that appears within 20 kilometers from the front line is basically being struck by an FPV drone or an artillery shell within seconds to minutes," Lange explains.

For a while, jammers that sever the radio links between drones and operators or that spoof GPS receivers were able to provide fairly reliable defense against human-controlled FPVs. But as autonomous navigation progressed, those electronic shields have gradually become less effective. Defenders must now contend with unjammable drones—ones that are attached to hair-thin optical fibers or that are capable of finding their way to their targets without external guidance. In this emerging struggle, the defenders' track records aren't very encouraging: The typical countermeasure is to try to shoot down the attacking drone with a service weapon. It's rarely successful.

"The attackers gain an immense advantage from unmanned systems," says Lange. "You can have a drone pop up from anywhere and it can wreak havoc. But from autonomy, they gain even more."

The self-navigating drones rely on image-recognition algorithms that have been around for over a decade, says Lange. And the mass deployments of drones on Ukrainian battlefields are enabling both Russian and Ukrainian technologists to create huge datasets that improve the training and precision of those AI algorithms.

While uncrewed aerial vehicles (UAVs) have received the most attention, the Ukrainian military is also deploying dozens of different kinds of drones on land and sea. Ukraine, struggling with the shortage of infantry personnel, began working on replacing a portion of human soldiers with wheeled ground robots in 2024. As of early 2026, thousands of ground robots are crawling across the gray zone along the front line in Eastern Ukraine. Most are used to deliver supplies to the front line or to help evacuate the wounded, but some "killer" ground robots fitted with turrets and remotely controlled machine guns have also been tested.

In mid-February, Ukrainian authorities released a video of a Ukrainian ground robot using its thermal camera to detect a Russian soldier in the dark of the night and then kill the invader with a round from a heavy machine gun. So far these robots are mostly controlled by a human operator, but the makers of these uncrewed ground vehicles say their systems are capable of basic autonomous operations, such as returning to base when radio connection is lost. The goal is to enable them to swarm so that one operator controls not one, but a whole herd of mesh-connected killer robots.

But Bryan Clark, senior fellow and director of the Center for Defense Concepts and Technology at the Hudson Institute, questions how quickly ground robots' abilities can progress. "Ground environments are very difficult to navigate in because of the terrain you have to address," he says. "The line of sight for the sensors on the ground vehicles is really constrained because of terrain, whereas an air vehicle can see everything around it."

To achieve autonomy, maritime drones, too, will require navigational approaches beyond AI-based image recognition, possibly based on star positions or electronic signals from radios and cell towers that are within reach, says Clark. Such technologies are still being developed or are in a relatively early operational stage.

> "The attackers gain an immense advantage from unmanned systems. You can have a drone pop up from anywhere and it can wreak havoc. But from autonomy, they gain even more."
>
> —MARC LANGE, DEFENSE ANALYST

Russia is not lagging behind. In fact, some analysts believe its autonomous systems may be slightly ahead of Ukraine's. For a good example of the Russian military's rapid evolution, they say, consider the long-range Iranian-designed Shahed drones. Since 2022, Russia has been using them to attack Ukrainian cities and other targets hundreds of kilometers from the front line.

"At the beginning, Shaheds just had a frame, a motor, and an inertial navigation system," Oleksii Solntsev, CEO of Ukrainian defense tech startup MaXon Systems, tells me. "They used to be imprecise and pretty stupid. But they are becoming more and more autonomous." Solntsev founded MaXon Systems in late 2024 to help protect Ukrainian civilians from the growing threat of Shahed raids.

First produced in Iran in the 2010s, Shaheds can carry 90-kilogram warheads up to 650 km (50-kg warheads can go twice as far). They cost around $35,000 per unit, compared to a couple of million dollars, at least, for a ballistic missile. The low cost allows Russia to manufacture Shaheds in high quantities, unleashing entire fleets onto Ukrainian cities and infrastructure almost every night.

The early Shaheds were able to reach a preprogrammed location based on satellite-navigation coordinates. Even one of these early models could frequently overcome the jamming of satellite-navigation signals with the help of an onboard inertial navigation unit. This was essentially a dead-reckoning system of accelerators and gyroscopes that estimate the drone's position from continual measurements of its motions.

Ukrainian defense forces learned to down Shaheds with heavy machine guns, but as Russia continued to innovate, the daily onslaughts started to become increasingly effective.

Today's Shaheds fly faster and higher, and therefore are more difficult to detect and take down. Between January 2024 and August 2025, the number of Shaheds and Shahed-type attack drones launched by Russia into Ukraine per month increased more than tenfold, from 334 to more than 4,000. In 2025, Ukraine found AI-enabling Nvidia chipsets in wreckages of Shaheds, as well as thermal-vision modules capable of locking onto targets at night.

"Now, they are interconnected, which allows them to exchange information with each other," Solntsev says. "They also have cameras that allow them to autonomously navigate to objects. Soon they will be able to tell each other to avoid a jammed

Ukrainian soldiers watch a GOR reconnaissance drone take to the sky near Pokrovsk in the Donetsk region, on 10 March 2025 [above]. A mobile electronic-warfare system built by PiranhaTech is demonstrated near Kyiv on 21 October 2025 [right].

TOP: ANDRIY DUBCHAK/FRONTLINER/GETTY IMAGES; BOTTOM: DANYLO ANTONIUK/ANADOLU/GETTY IMAGES

region or an area where one of them got intercepted."

These Russian-manufactured Shaheds, which Russian forces call Geran-2s, are thought to be more capable than the garden variety Shahed-136s that Iran has lately been launching against targets throughout the Middle East. Even the relatively primitive Shahed-136s have done considerable damage, according to press accounts.

Those Shahed successes may accrue, at least in part, from the fact that the United States and Israel lack Ukraine's long experience with fending them off. In just two days in early March, upward of a thousand drones, mostly Shaheds, were launched against U.S. and Israeli targets, with hundreds of them reportedly finding their marks.

One attack, caught on videotape, shows a Shahed destroying a radar dome at the U.S. navy base in Manama, Bahrain. U.S. forces were understood to be attempting to fend off the drones by striking launch platforms, dispatching fighter aircraft to shoot them down, and by using some extremely costly air-defense interceptors, including ones meant to down ballistic missiles. On 4 March, CNN reported that in a congressional briefing the day before, top U.S. defense officials, including Secretary of Defense Pete Hegseth, acknowledged that U.S. air defenses weren't keeping up with the onslaught of Shahed drones.

Russia is also starting to field a newer generation of attack drones. One of these, the V2U, has been used to strike targets in the Sumy region of northeastern Ukraine. The V2U drones are outfitted with Nvidia Jetson Orin processors and run computer-vision software and AI algorithms that allow the drones to navigate even where satellite navigation is jammed.

The sale of Nvidia chips to Russia is banned under U.S. sanctions against the country. However, press reports suggest that the chips are getting to Russia via intermediaries in India.

MaXon Systems is one of several companies working to fend off the nightly drone onslaught. Within one year, the company developed and battle-tested a Shahed interception system that hints at the sci-fi future envisioned by Azhnyuk. For a system to be capable of reliably defending against autonomous weaponry, it, too, needs to be autonomous.

MaXon's solution consists of ground turrets scanning the sky with infrared sensors, with additional input from a network of radars that detects approaching Shahed drones at distances of, typically, 12 to 16 km. The turrets fire autonomous fixed-winged interceptor drones, fitted with explosive warheads, toward the approaching Shaheds at speeds of nearly 300 km/h. To boost the chances of successful interception, MaXon is also fielding an airborne anti-Shahed fortification system consisting of helium-filled aerostats hovering above the city that dispatch the interceptors from a higher altitude.

"We are trying to increase the level of automation of the system compared to existing solutions," says Solntsev. "We need automatic detection, automatic takeoff, and automatic mid-track guidance so that we can guide the interceptor before it can itself flock the target."

In November 2025, the Ukrainian military announced it had been conducting successful trials of the Merops Shahed drone interceptor system developed by the U.S. startup Project Eagle, another of former Google CEO Eric Schmidt's Ukraine defense ventures. Like the MaXon gear, the system can operate largely autonomously and has so far downed over 1,000 Shaheds.

Despite the progress on both sides, analysts say that the kind of robotic warfare imagined by Azhnyuk won't be a reality for years.

"The software for drone collaboration is there," says Kate Bondar, a former policy advisor for the Ukrainian government and currently a research fellow at the U.S. Center for Strategic and International Studies. "Drones can fly in labs, but in real life, [the forces] are afraid to deploy them because the risk of a mistake is too high," she adds.

In Bondar's view, powerful AI-equipped drones won't be deployed in large numbers given the current prices for high-end processors and other advanced components. And, she adds, the more autonomous the system needs to be, the more expensive are the processors and sensors it must have. "For these cheap attack drones that fly only once, you don't install a high-resolution camera that [has] the resolution for AI to see properly," she says. "[You install] the cheapest camera. You don't want expensive chips that can run AI algorithms either. Until we can achieve this balance of technological sophistication, when a system can conduct a mission but at the lowest price possible, it won't be deployed en masse."

While existing AI systems are doing a good job recognizing and following large objects like Shaheds or tanks, experts question their ability to reliably distinguish and pursue smaller and more nimble or inconspicuous targets. "When we're getting into more specific questions, like can it distinguish a Russian soldier from a Ukrainian soldier or at least a soldier from a civilian? The answer is no," says Bondar. "Also, it's one thing to track a tank, and it's another to track infantrymen riding buggies and motorcycles that are moving very fast. That's really challenging for AI to track and strike precisely."

Clark, at the Hudson Institute, says that although the AI algorithms used to guide the Russian and Ukrainian drones are "pretty good," they rely on information provided by sensors that "aren't good enough."

"You need multiphenomenology sensors that are able to look at infrared and visual and, in some cases, different parts of the infrared spectrum to be able to figure out if something is a decoy or real target," he says.

German defense analyst Lange agrees that right now, battlefield AI image-recognition systems are too easily fooled. "If you compress reality into a 2D

Department of Electrical and Computer Engineering
Graduate School of Engineering and Management
Air Force Institute of Technology (AFIT) | Dayton, Ohio

The Department of Electrical and Computer Engineering at AFIT is seeking applications for tenured or tenure-track faculty positions. Applicants must have earned a doctorate in Computer Science, Computer Engineering, Electrical Engineering, or a closely related discipline by the time of appointment (anticipated 1 September 2026).

Applicants with expertise in the following areas are of particular interest: Artificial Intelligence, supervised and unsupervised machine learning, neural networks, neuromorphic computing, deep learning, reinforcement learning, natural language processing, generative AI, explainable AI, human-AI interaction; Autonomy, agents, multi-agent systems, and robotics; Quantum computing, quantum communications, quantum sensing, quantum materials, and quantum precision navigation and timing; Field Programmable Gate Arrays (FPGA) design, optimization, and applications, or Very Large-Scale Integration (VLSI) design and development; cybersecurity; Electronic Warfare (EW), adaptive or cognitive EW.

The Department of Electrical and Computer Engineering offers accredited M.S. and Ph.D. programs in Electrical Engineering, Computer Engineering, and Computer Science, and an M.S. program in Cyber Operations.

Include:
- Letter of application with the USAJOBS ad number
- Resume (maximum two pages) • Curriculum vitae (no photographs)
- Transcripts for all listed degrees on curriculum vitae (official copies required later)
- Statement of research plans and graduate-level teaching philosophy (one page each)
- List of three professional references including name, complete mailing address, email address, phone number

For more information & to apply, visit:

Applicants must be U.S. citizens and hold or be able to obtain a security clearance.
For additional information on AFIT and the Department of Electrical and Computer Engineering, visit **https://www.afit.edu/ENG/**. The USAJOBS application closes **06/30/2026**.

One of the most influential reference resources for engineers around the world.

For over 100 years, *Proceedings of the IEEE* has been the leading journal for engineers looking for in-depth tutorial, survey, and review coverage of the technical developments that shape our world.

To learn more and start your subscription today, visit
ieee.org/proceedings-subscribe

IEEE

image, a lot of things can be easily camouflaged—like what Russia did recently, when they started drawing birds on the back of their drones," he says.

T**To make Ukraine's** emerging uncrewed ground vehicles (UGVs) equally self-sufficient will be an even greater task, in Clark's view. Still, Bondar expects major advances to materialize within the next several years, even if humans are still going to be part of the decision-making loop.

"I think in two or three years, we will have pretty good full autonomy, at least in good weather conditions," she says, referring to aerial drones in particular. "Humans will still be in the loop for some years, simply because there are so many unpredictable situations when you need an intervention. We won't be able to fully rely on the machine for at least another 10 or 15 years."

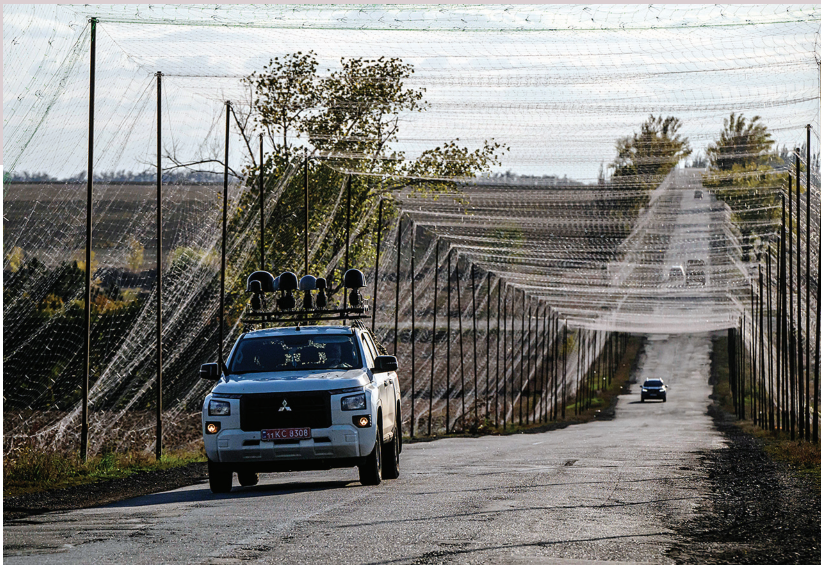
Ukrainian defenders are apprehensive about that autonomous future. The boom of drone innovation has come hand in hand with the development of sophisticated jamming and radio-frequency detection systems. But a lot of that innovation will become obsolete once the pendulum swings away from human control. Ukrainians got their first taste of dealing with unjammable drones in mid-2024, when Russia began rolling out fiber-optic tethered drones. Now they have to brace for a threat on a much larger scale.

"Today, we have a situation where we have lots of signals on the battlefield, but in the near future, in maybe two to five years, UAVs are not going to be sending any signals," says Oleksandr Barabash, CTO of Falcons, a Ukrainian startup that has developed a smart radio-frequency detection system capable of revealing precise locations of enemy radio sources such as drones, control stations, and jammers.

Last September, Falcons secured funding from the U.S.-based dual-use tech fund Green Flag Ventures to scale production of its technology and work toward NATO certification. But Barabash admits that its system, like all technologies fielded in Ukrainian war zones, has an expiration date. Instead of radio-frequency detectors, Barabash thinks, the next R&D push needs to focus on passive radar systems capable of identifying small and fast-moving targets based on the signal from sources like TV towers or radio transmitters that propagate through the environment and are reflected by those moving targets. Passive radars have a significant advantage in the war zone, according to Barabash. Since they don't emit their own signal, they can't be that easily discovered by the enemy.

"Active radar is emitting signals, so if you are using active radars, you are target No. 1 on the front line," Barabash says.

Bondar, on the other hand, thinks that the increased onboard compute power needed for AI-controlled drones will, by itself, generate enough electromagnetic radiation to prevent autonomous drones from ever operating completely undetectably.



A truck outfitted with signal-jamming gear drives under antidrone nets near Oleksandriya, in eastern Ukraine, on 2 October 2025.

ED JONES/AFP/GETTY IMAGES

"You can have full autonomy, but you will still have systems onboard that emit electromagnetic radiation or heat that can be detected," says Bondar. "Batteries emit electromagnetic radiation, motors emit heat, and [that heat can be] visible in infrared from far away. You just need to have the right sensors to be able to identify it in advance." She adds that that takeaway is "how capable contemporary detection systems have become and how technically challenging it is to design drones that can reliably operate in the Ukrainian battlefield environment."

W**When autonomous drones** become a standard weapon of war, their threat will extend far beyond the battlefields of Ukraine. Autonomous turrets and drone-interceptor fortification might soon dot the perimeter of European cities, particularly in the eastern part of the continent. Nefarious actors from all over the world have closely watched Ukraine and taken notes, warns Lange. Today, FPV drones are being used by Islamic terrorists in Africa and Mexican drug cartels to fight against local authorities.

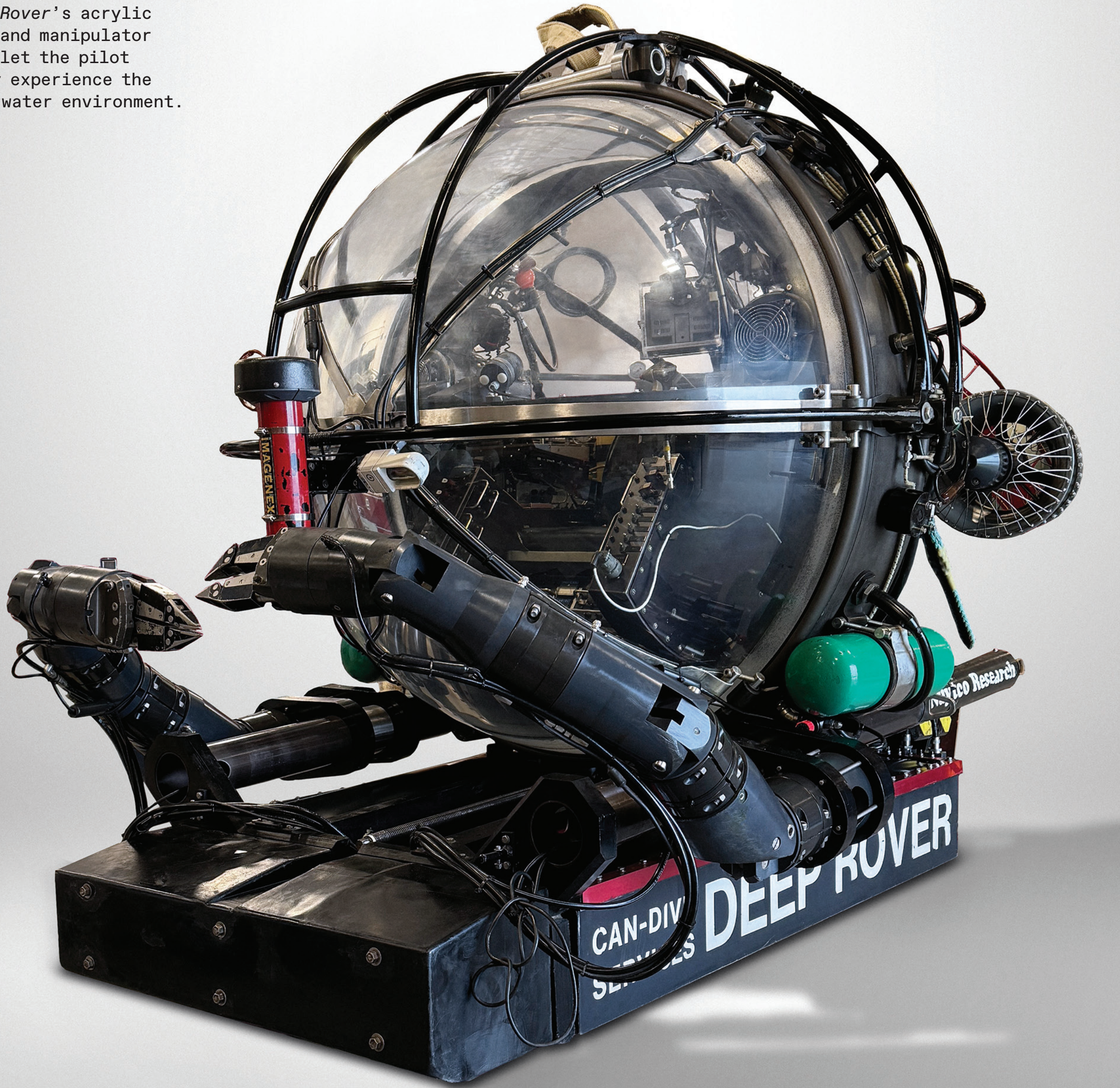
When autonomous killing machines become widely available, it's likely that no city will be safe. "We might see nets above city centers, protecting civilian streets," Lange says. "In every case, the West needs to start performing similar kinetic-defense development that we see in Ukraine. Very rapid iteration and testing cycles to find solutions."

Azhnyuk is concerned that the historic defenders of Europe—the United States and the European countries themselves—are falling behind. "We are in danger," he says. While Russia and Ukraine made major strides in their drones and countermeasures over the past year, "Europe and the United States have progressed, in the best-case scenario, from the winter-of-2022 technology to the summer-of-2022 technology.

"The gap is getting wider," he warns. "I think the next few years are very dangerous for the security of Europe." ■

BY ALLISON MARSH

Past Forward



Deep Rover's acrylic hull and manipulator arms let the pilot truly experience the underwater environment.

INGENIUM – CANADA'S MUSEUMS OF SCIENCE AND INNOVATION

All Alone in the Abyss

***Deep Rover* was** a one-person submersible designed to satisfy human curiosity about the underwater world. As the rover moved freely through the water down to depths of 1,000 meters, its operator sat in relative comfort in the cab, inside a clear acrylic bubble with panoramic views—an inverted fishbowl, with the human immersed in breathable air while the sea creatures looked in.

The team behind *Deep Rover* included Canadian inventor and entrepreneur Phil Nuytten, U.S. marine biologist Sylvia Earle, and British marine engineer and submarine designer Graham Hawkes. The sub was originally intended to aid in offshore oil exploration and drilling. But when oil markets collapsed shortly after *Deep Rover*'s launch in 1984, its mission broadened to scientific research. It conducted some 280 dives over its eight-year lifetime, and a prototype was featured in the 1989 deep-sea sci-fi movie *The Abyss*. ■

FOR MORE ON DEEP ROVER, see spectrum.ieee.org/pastforward-apr2026

IEEE *Access*®

Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Harness the publishing power of IEEE *Access*®.

IEEE *Access* is a multidisciplinary open access journal offering high-quality peer review, with an expedited, binary review process of 4 to 6 weeks. As a journal published by IEEE, IEEE *Access* offers a trusted solution for authors like you to gain maximum exposure for your important research.

Explore the many benefits of IEEE *Access*:

- **Receive high-quality, rigorous peer review** in only 4 to 6 weeks
- **Reach millions of global users** through the IEEE *Xplore*® digital library by publishing open access
- **Submit multidisciplinary articles** that may not fit in narrowly focused journals
- **Obtain detailed feedback** on your research from highly experienced editors
- **Establish yourself as an industry pioneer** by contributing to trending, interdisciplinary topics in one of the many topical sections IEEE *Access* hosts
- **Present your research to the world quickly** since technological advancement is ever-changing
- **Take advantage of features** such as multimedia integration, usage and citation tracking, and more
- **Publish without a page limit** for $1,750 per article

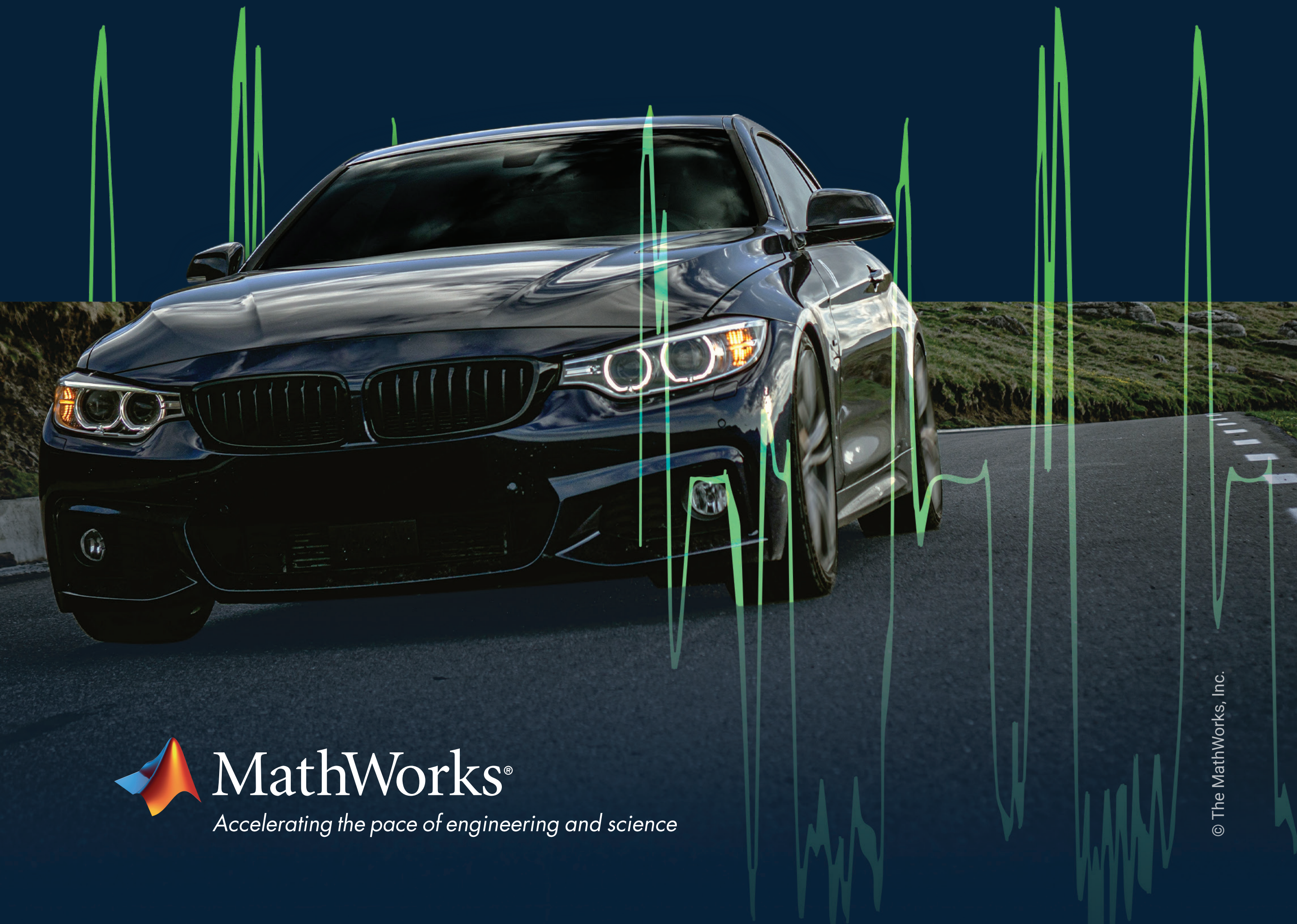
⋮ Learn more at ieeeaccess.ieee.org

IEEE

MATLAB
FOR AI

Boost system design and simulation with explainable and scalable AI. With MATLAB and Simulink, you can easily train and deploy AI models.

mathworks.com/ai



MathWorks®
Accelerating the pace of engineering and science

© The MathWorks, Inc.